Massachusetts Institute of Technology
Department of Economics
Working Paper Series

ℓ_1-PENALIZED QUANTILE REGRESSION IN HIGH-DIMENSIONAL SPARSE MODELS

Alexandre Belloni
Victor Chernozhukov

Working Paper 09-11
April 24, 2009

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

ℓ_1-PENALIZED QUANTILE REGRESSION IN HIGH-DIMENSIONAL SPARSE MODELS

BY ALEXANDRE BELLONI AND VICTOR CHERNOZHUKOV[*,†]

Duke University and Massachusetts Institute of Technology

We consider median regression and, more generally, quantile regression in high-dimensional sparse models. In these models the overall number of regressors p is very large, possibly larger than the sample size n, but only s of these regressors have non-zero impact on the conditional quantile of the response variable, where s grows slower than n. Since in this case the ordinary quantile regression is not consistent, we consider quantile regression penalized by the ℓ_1-norm of coefficients (ℓ_1-QR). First, we show that ℓ_1-QR is consistent at the rate $\sqrt{s/n}\sqrt{\log p}$, which is close to the oracle rate $\sqrt{s/n}$, achievable when the minimal true model is known. The overall number of regressors p affects the rate only through the $\log p$ factor, thus allowing nearly exponential growth in the number of zero-impact regressors. The rate result holds under relatively weak conditions, requiring that s/n converges to zero at a super-logarithmic speed and that regularization parameter satisfies certain theoretical constraints. Second, we propose a pivotal, data-driven choice of the regularization parameter and show that it satisfies these theoretical constraints. Third, we show that ℓ_1-QR correctly selects the true minimal model as a valid submodel, when the non-zero coefficients of the true model are well separated from zero. We also show that the number of non-zero coefficients in ℓ_1-QR is of same stochastic order as s, the number of non-zero coefficients in the minimal true model. Fourth, we analyze the rate of convergence of a two-step estimator that applies ordinary quantile regression to the selected model. Fifth, we evaluate the performance of ℓ_1-QR in a Monte-Carlo experiment, and provide an application to the analysis of the international economic growth.

[*] First version: December, 2007, This version: April 19, 2009.

[†] The authors gratefully acknowledge the research support from the National Science Foundation.

AMS 2000 subject classifications: Primary 62H12, 62J99; secondary 62J07

Keywords and phrases: median regression, quantile regression, sparse models

1. Introduction. Quantile regression is an important statistical method for analyzing the impact of regressors on the conditional distribution of a response variable (cf. Laplace [22], Koenker and Bassett [20]). In particular, it captures the heterogeneity of the impact of regressors on the different parts of the distribution [7], exhibits robustness to outliers [19], has excellent computational properties [29], and has a wide applicability [19]. The asymptotic theory for quantile regression is well-developed under both fixed number of regressors and increasing number of regressors. The asymptotic theory under fixed number of regressors is given by Koenker and Bassett [20], Portnoy [28], Gutenbrunner and Jurečková [14], Knight [17], Chernozhukov [9] and others. The asymptotic theory under increasing number of regressors is given in He and Shao [15] and Belloni and Chernozhukov [4, 5], covering the case where the number of regressors p is negligible relative to the sample size n ($p = o(n)$).

In this paper, we consider quantile regression in high-dimensional sparse models (HDSMs). In such models, the overall number of regressors p is very large, possibly much larger than the sample size n. However, the number s of significant regressors – those having a non-zero impact on the response variable – is smaller than the sample size, that is, $s = o(n)$. The HDSMs ([8], [26]) have emerged to deal with many new applications, arising in biometrics, signal processing, machine learning, econometrics, and other areas of data analysis, where high-dimensional data sets have become widely available.

A number of papers began to investigate estimation of HDSMs, primarily focusing on penalized mean regression, with ℓ_1-norm acting as a penalty function. Candes and Tao [8] demonstrated that, remarkably, an estimator, called the Dantzig selector, achieves the rate $\sqrt{s/n}\sqrt{\log p}$, which is very close to the oracle rate $\sqrt{s/n}$ obtainable when the significant regressors are known. Thus the estimator can be consistent even under very rapid, nearly exponential growth in the total number of regressors p. Meinshausen and Yu [26] and Zhang and Huang [39] demonstrated similar striking results for the ℓ_1-penalized least squares proposed by Tibshirani [35]. van der Geer [37] derived valuable finite sample bounds on empirical risk for ℓ_1-penalized estimators in generalized linear models. Fan and Lv [11] used screening and derived asymptotic results under even weaker conditions on p. There were many other interesting developments, which we shall not review here.

Our paper's contribution is to develop, within the HDSM framework, a set of results on model selection and rates of convergence for quantile regression. Since ordinary quantile regression is not consistent in HDSMs, we consider quantile regression penalized by the ℓ_1-norm of parameter coefficients. We show that the ℓ_1-penalized quantile regression is consistent at the rate $\sqrt{s/n}\sqrt{\log p}$, which is close to the oracle rate $\sqrt{s/n}$ achievable when the true minimal model is known. In order to make the penalized estimator practical, we propose a pivotal, data-driven choice of the regularization parameter, and show that this choice leads to the same sharp convergence rate. Further, we show that the penalized quantile regression correctly selects the true minimal model as a valid submodel, when the non-zero coefficients of the true model are well separated from zero. We also analyze a two-step estimator that applies standard quantile regression to the selected model and aims at reducing the bias of the penalized quantile estimator. We illustrate the use of the penalized and post-penalized estimators with a Monte carlo experiment and an international economic growth example. Thus, our results contribute to the literature on HDSMs by examining a new class of problems. Moreover, our proof strategy, developed to cope with non-linearities and non-smoothness of quantile regression, may be of interest in other M-estimation problems. (We provide more detailed comparisons to the literature in Section 2.)

Finally, let us comment on the role of computational considerations in our analysis. The choice of the ℓ_1-penalty function arises from considering a tradeoff between statistical efficiency and computational efficiency, with the latter being of particular importance in high-dimensional applications. Indeed, in model selection problems, the statistical efficiency criterion favors the use of the ℓ_0-penalty functions (Akaike [1] and Schwarz [32]), where the ℓ_0-penalty counts the the number of non-zero components of a parameter vector. However, the computational efficiency criterion favors the use of convex penalty functions. Indeed, convex penalty functions lead to efficient convex programming problems ([27]); in contrast, the ℓ_0-penalty functions lead to inefficient combinatorial search problems, plagued by the computational curse of dimensionality. Precisely because it is a convex function that is closest to the ℓ_0-penalty (e.g. [30]), the ℓ_1-penalty has emerged to play a central role in HDSMs, in general (e.g. [25]), and in our analysis, in particular. In other words, the use of the ℓ_1-penalty takes us close to performing the most effective model selection, while respecting the computational efficiency constraint.

We organize the rest of the paper as follows. In Section 2, we introduce the problem and some simple primitive assumptions D.1-D.4, and propose pivotal choices for the regularization parameter. We also describe our key results under D.1-D.4, and provide detailed comparisons with the literature. In Section 3, we develop the main results under conditions E.1-E.5, which are implied by D.1-D.4, and also hold much more generally. Section 4 analysis the pivotal choice of the penalization parameter. In Section 5, we carry out a computational experiment and provide an application to an international growth example. In Section 6, we provide conclusions and discuss possible extensions. In Appendix A, we verify that conditions E.1-E.5 are implied by conditions D.1-D.4 and also hold more generally.

1.1. *Notation.* In what follows, we implicitly index all parameter values by the sample size n, but we omit the index whenever this does not cause confusion. We carry out all of the asymptotic analysis as $n \to \infty$. We use the notation $a \lesssim b$ to denote that $a = O(b)$, that is $a \leq cb$ for all sufficiently large n, for some constant $c > 0$ that does not depend on n, and we use $a \lesssim_p b$ to denote that $a = O_p(b)$; we use $a \simeq b$ to denote $a \lesssim b \lesssim a$ and $a \simeq_p b$ to denote $a \lesssim_p b \lesssim_p a$. We also use the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We denote ℓ_2-norm by $\|\cdot\|$, ℓ_1-norm by $\|\cdot\|_1$, ℓ_∞-norm by $\|\cdot\|_\infty$, and the ℓ_0-"norm" by $\|\cdot\|_0$.

2. Basic Settings, the Estimator, and Overview of Results. In this section, we formulate the setting, the estimator, and state primitive regularity conditions. We also provide an overview of the main results.

2.1. *Basic Setting.* The set-up of interest corresponds to a parametric quantile regression model, where the dimension p of the underlying model increases with the sample size n. Namely, we consider a response variable y and p-dimensional covariates x such that the u-th conditional quantile function of y given x is given by

$$Q_{y|x}(u) = x'\beta(u), \quad \beta(u) \in \mathbb{R}^p. \tag{2.1}$$

We consider the case where the dimension p of the model is large, possibly much larger than the available sample size n, but the true model $\beta(u)$ is sparse having only $s = s(u) < p$ non-zero components. Throughout the paper the quantile index $u \in (0, 1)$ is fixed.

The population coefficient $\beta(u)$ is known to be a minimizer of the criterion function

$$Q_u(\beta) = \mathrm{E}[\rho_u(y - x'\beta)], \tag{2.2}$$

where $\rho_u(t) = (u - 1\{t \leq 0\})t$ is the asymmetric absolute deviation function [20]. Given a random sample $(y_1, x_1), \dots, (y_n, x_n)$, the quantile regression estimator $\widehat{\beta}(u)$ of $\beta(u)$ is defined as a minimizer of

$$\widehat{Q}_u(\beta) = \mathbb{E}_n\left[\rho_u(y_i - x_i'\beta)\right] \tag{2.3}$$

where $\mathbb{E}_n\left[f(y_i, x_i)\right] := n^{-1}\sum_{i=1}^n f(y_i, x_i)$ denotes the empirical expectation of a function f in the given sample.

In the high-dimensional settings, particularly when $p \geq n$, quantile regression is generally not consistent, which motivates the use of penalization in order to remove all or at least nearly all regressors whose population coefficients are zero, thereby possibly restoring consistency. The penalization that has been proven to be quite useful in least squares settings is the ℓ_1-penalty leading to the lasso estimator [35].

2.2. *The Choice of Estimator, Linear Programming Formulation, and Its Dual.* The ℓ_1-penalized quantile regression estimator $\widehat{\beta}(u)$ is a solution to the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \widehat{Q}_u(\beta) + \frac{\lambda}{n} \sum_{j=1}^{p} |\beta_j|. \tag{2.4}$$

When the solution is not unique, we define $\widehat{\beta}(u)$ as a basic solution having the minimal number of non-zero components. The criterion function in (2.4) is the sum of the criterion function (2.3) and a penalty function given by a scaled ℓ_1-norm of the parameter vector. This ℓ_1-penalized quantile regression or quantile regression lasso has been considered by Knight and Fu [18] under the small (fixed) p asymptotics.

For computational purposes, it is important to note that the penalized quantile regression problem (2.4) is equivalent to the following linear programming problem

$$\begin{array}{ll} \displaystyle\min_{\xi^+,\xi^-,\beta^+,\beta^- \in \mathbb{R}_+^{2n+2p}} & \mathbb{E}_n\left[u\xi_i^+ + (1-u)\xi_i^-\right] + \frac{\lambda}{n}\sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) \\ & \xi_i^+ - \xi_i^- = y_i - x_i'(\beta^+ - \beta^-), \quad i = 1, \ldots, n. \end{array} \tag{2.5}$$

The problem minimizes a sum of ℓ_1-norm of the absolute positive β_j^+ and negative β_j^- parts of the parameter $\beta_j = \beta_j^+ - \beta_j^-$ and of an average of asymmetrically weighted residuals ξ_i^+ and ξ_i^-. The linear programming formulation (2.5) is useful for computation of the estimator, particularly in high-dimensional applications. There are a number of efficient, that is, polynomial time, algorithms for the linear programming problem (2.5). Using these algorithms, one can compute the estimator (2.4) efficiently, avoiding the computational curse of dimensionality.

Furthermore, for both computational and theoretical purposes, it is important to note that the primal problem (2.5) has the following dual problem:

$$\begin{array}{ll} \displaystyle\max_{a \in \mathbb{R}^n} & \mathbb{E}_n\left[y_i a_i\right] \\ & |\mathbb{E}_n\left[x_{ij} a_i\right]| \leq \frac{\lambda}{n}, \quad j = 1, \ldots, p, \\ & (u-1) \leq a_i \leq u, \quad i = 1, \ldots, n. \end{array} \tag{2.6}$$

The dual problem maximizes the correlation between the response variable and the rank scores subject to the condition requiring the rank scores to be approximately uncorrelated with the regressors. This condition is reasonable, since the true rank scores, defined as $a_i^*(u) = (u - 1\{y_i \leq x_i'\beta(u)\})$, should be independent of regressors x_i. This follows because by (2.1) the event $\{y_i \leq x_i'\beta(u)\}$ is equivalent to the event $\{u_i \leq u\}$, for a standard uniformly distributed variable u_i which is independent of x_i.

Since both primal and dual problems are feasible, by strong duality for linear programming the optimal values of (2.6) equals the optimal value of (2.4) (see, for example, Bertsimas and Tsitsiklis [6]). The optimal solution to the dual problem plays an important role in our analysis, helping us control the sparseness of the penalized estimator $\widehat{\beta}(u)$ as well as choose the penalization parameter λ. Of course, the optimal solution to the dual problem (2.6) also plays an important role in the non-penalized case, with $\lambda = 0$, yielding the regression generalization of Hajek-Sidak rank scores (Gutenbrunner and Jurečková [14]).

Another potential approach worth considering is the Dantzig selector approach of Candes and Tao [8], proposed in the context of mean regression. We can extend this approach to quantile regression by defining the estimator as solution to the following problem:

$$\inf_{\beta \in \mathbb{R}^p} \sum_{j=1}^{p} |\beta_j| : \quad \|\widehat{S}_u(\beta)\|_\infty \leq \frac{\gamma}{n}, \tag{2.7}$$

where λ is a penalization parameter, and $\widehat{S}_u$ is a subgradient of the quantile regression objective function $\widehat{Q}_u(\beta)$:

$$\widehat{S}_u(\beta) = \mathbb{E}_n[(1\{y_i \leq x_i'\beta\} - u)x_i]. \tag{2.8}$$

The estimator (2.7) minimizes the ℓ_1-norm of the coefficients subject to a goodness-of-fit constraint.

On computational grounds, we prefer the ℓ_1-penalized estimator (2.4) over to the Dantzig selector estimator (2.7). The reason is that the subgradient $\widehat{S}_u$ in (2.8) is a piece-wise constant function in parameters, leading to a serious difficulty in computing the estimator (2.7). In particular, the problem (2.7) can be recast as a mixed integer programming problem with n binary variables, for which (generally) there is no known polynomial time algorithm. (In sharp contrast, in the mean regression case the subgradient is a linear function, $\widehat{S}(\beta) = \mathbb{E}_n[(y_i - x_i\beta)x_i]$, corresponding to the objective function $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i\beta)^2]/2$. Accordingly, in the mean regression case, the optimization problem can be recast as a linear programming problem, for which there are polynomial time algorithms.)

Another idea for formulating a Dantzig type estimator for quantile regression would be to minimize the ℓ_1 norm of the coefficients subject to a convex goodness-of-fit constraint, namely

$$\min_{\beta \in \mathbb{R}^p} \sum_{j=1}^{p} |\beta_j| \ : \widehat{Q}_u(\beta) \leq \gamma. \tag{2.9}$$

Since the constraint set $\{\beta : \widehat{Q}_u(\beta) \leq \gamma\}$ is piece-wise linear and convex, this problem is equivalent to a linear programming problem. Of course, this is hardly a surprise, since this problem is equivalent to an ℓ_1-penalized quantile regression problem (2.4) that we started with in the first place. Indeed, for every feasible choice of γ in (2.9) there is a feasible choice of λ that makes the solutions to (2.9) and to (2.4) identical. To see this, fix a γ and let κ denote the optimal value of the Lagrange multiplier for the constraint $\widehat{Q}_u(\beta) \leq \gamma$, then the problem (2.9) is equivalent to $\min_{\beta \in \mathbb{R}^p} \sum_{j=1}^{p} |\beta_j| + \kappa(\widehat{Q}_u(\beta) - \gamma)$, which is then equivalent to the original problem (2.4) with $\lambda = n/\kappa$. Therefore it suffices to focus our analysis on the original problem.

2.3. *The Choice of the Regularization Parameter.* Here we propose a pivotal, data-driven choice for the regularization parameter value λ. We shall verify in Section 4 that such choice will agree with our theoretical choice of λ maximizing the speed of convergence of the penalized estimate to the true parameter value.

Because the objective function in the primal problem (2.4) is not pivotal in either small or large samples, finding a pivotal λ appears to be difficult a priori. However, instead of looking at the primal problem, let us look at its linear programming dual (2.6), which requires that

$$|\mathbb{E}_n [x_{ij} a_i]| \leq \frac{\lambda}{n}, \text{for all } j = 1, \ldots, p \Leftrightarrow \|\mathbb{E}_n [x_i a_i]\|_\infty \leq \frac{\lambda}{n}. \tag{2.10}$$

This restriction requires that potential rank scores must be approximately uncorrelated with regressors. It then makes sense to select λ so that the true rank scores

$$a_i^*(u) = (u - 1\{y_i \leq x_i'\beta(u)\}) \ \text{ for } i = 1, \ldots, n$$

satisfy this constraint. That is, we can potentially set $\lambda = \Lambda_n$, where

$$\Lambda_n = n \, \|\mathbb{E}_n [x_i a_i^*(u)]\|_\infty . \tag{2.11}$$

Of course, since we do not observe the true rank scores, this choice is not available to us. The key observation is that the finite sample distribution of Λ_n is pivotal conditional on the regressors $x_1, \ldots, x_n$. We know that rank scores can be represented almost surely as

$$a_i^*(u) = (u - 1\{u_i \leq u\}), \quad \text{for } i = 1, \ldots, n,$$

where $u_1, \ldots, u_n$ are i.i.d. uniform $(0, 1)$ random variables, independently distributed from the regressors, $x_1, \ldots, x_n$. Thus, we have

$$\Lambda_n = n \left\| \mathbb{E}_n \left[x_i (u - 1\{u_i \leq u\}) \right] \right\|_\infty, \tag{2.12}$$

which has a known distribution conditional on $x_1, \ldots, x_n$. Therefore we can use the tail quantiles Λ_n as our choice for λ. In particular, we set $\lambda = \lambda(x_1, \ldots, x_n)$ as the $1 - \alpha_n$ quantile of Λ_n

$$\lambda = \inf\{c : P(\Lambda_n \leq c | x_1, \ldots, x_n) \geq 1 - \alpha_n\}, \tag{2.13}$$

where $\alpha_n \searrow 0$ at some rate to be determined below.

Finally, let us note that we can also derive the pivotal quantity Λ_n, and thus also our choice of the regularization parameter λ, from the subgradient characterization of optimality for the primal problem (2.4).

2.4. *Primitive Conditions.* We follow Huber's framework of high-dimensional parameters [16], which formally consists of a sequence of models with parameter dimension $p = p_n$ tending to infinity as the sample size n grows to infinity. Thus, the parameters of the models, the parameter space, and the parameter dimension are all indexed by the sample size n. However, following Huber's convention, we will omit the index n whenever this does not cause confusion. Let us consider the following set of conditions:

D.1. Sampling. Data $(y_i, x_i')', i = 1, \ldots, n$ are an i.i.d. sequence of real $(1 + p)$-vectors, with the conditional u-quantile function given by (2.1), and with the first component of x_i equal to one.

D.2. Sparseness of the True Model. The number of non-zero components of $\beta(u)$ is bounded by $1 \leq s = s_n \leq n / \log(n \vee p)$.

D.3. Smooth Conditional Density. The conditional density $f_{y_i|x_i}(y|x)$ and its derivative $\frac{\partial}{\partial y} f_{y_i|x_i}(y|x)$ are bounded above uniformly in y and x ranging over supports of y_i and x_i, and uniformly in n.

D.4. Identifiability in Population and Well-Behaved Regressors. Eigenvalues of the population design matrix $\mathrm{E}[x_i x_i']$ are bounded above and away from zero, and $\sup_{\|\alpha\|=1} \mathrm{E}[|x_i'\alpha|^3]$

is bounded above, uniformly in n. The conditional density evaluated at the conditional quantile, $f_{y_i|x_i}(x'\beta(u)|x)$ is bounded away from zero, uniformly in x ranging over the support of x_i, and uniformly in n.

The conditions D.1-D.4 stated above are a set of simple conditions that ensure that the high-level conditions developed in Section 3 hold. These conditions allow us to demonstrate the general applicability of our results and straightforwardly compare to other results in the literature. In particular, condition D.1 imposes random sampling on the data, which is a conventional assumption in asymptotic statistics (e.g [38]). Condition D.2 requires that the effective dimension of the true model is smaller than the sample size. Condition D.3 imposes some smoothness on the conditional distribution of the response variable. Condition D.4 requires the population design matrix to be uniformly non-singular and the regressors' moments to be well-behaved.

Further, let $\phi(k)$ be the maximal k-sparse eigenvalue of the empirical design matrix $\mathbb{E}_n\left[x_i x_i'\right]$, that is,

$$\phi(k) = \sup_{\|\alpha\|\leq 1, \|\alpha\|_0 \leq k} \mathbb{E}_n\left[(\alpha' x_i)^2\right]. \tag{2.14}$$

Following Meinshausen and Yu [26], we will state our general results on convergence rates of the penalized estimator in terms of the maximal sparse eigenvalue $\phi(m_0)$. Meinshausen and Yu [26] worked with $m_0 = n \wedge p$ as an initial upper bound on the zero norm of the penalized estimator. In this paper we can work with a smaller m_0, in particular, under D.1-D.4, we can work with

$$m_0 = p \wedge (n/\log(n \vee p)),$$

as this provides a valid initial bound on the zero norm of our penalized estimator under a suitable choice of the penalization parameter.

By using an assumption on the growth rate of $\phi(k)$, we avoid imposing Candes and Tao's [8] uniform uncertainty principle on the empirical design matrix $\mathbb{E}_n\left[x_i x_i'\right]$. Meinshausen and Yu [26] argue that the assumption in terms of $\phi(k)$ are less stringent than the uniform uncertainty principle, since it allows for non-vanishing correlation between the regressors. Meinshausen and Yu [26] provide a thorough discussion of the behavior of $\phi(n)$ in many cases of interest. In particular, they show that the condition $\phi(n) \lesssim_p 1$ appears reasonable in several cases (for example, when the empirical design matrix is block diagonal). Note that if the intercept is included as a covariate we have $\phi(1) \geq 1$. For the purposes of a basic overview of results in the next subsection, we employ the assumption

$$\phi(n/\log(n \vee p)) \lesssim_P 1. \tag{2.15}$$

which will cover standard Gaussian regressors and some other regressors considered in Meinshausen and Yu [26] (because $\phi(n/\log(n \vee p)) \leq \phi(n)$). Furthermore, in our general analysis presented in Section 3, we do not impose (2.15) and allow for the sparse eigenvalue $\phi(n/\log(n \vee p))$ to diverge, which should permit for situations with regressors having tails thicker than Gaussian.

In order to illustrate our conditions we employ the following canonical examples throughout the paper.

Example 1 (Isotropic Normal Design). Let us consider estimating the median ($u = 1/2$) of the following regression model

$$y = x'\beta_0 + \varepsilon,$$

where the covariate $x_1 = 1$ is the intercept and the covariates $x_{-1} \sim N(0, I)$, and the errors are independent identically distributed with a smooth probability density function which is positive at zero and has bounded derivatives. This example satisfies conditions D.1, D.3, D.4, and D.2 if $\|\beta_0\|_0 \leq s = o(n/\log(n \vee p))$. Moreover, the maximal k-sparse eigenvalues for $k \leq n$ satisfy

$$\phi(k) := \sup_{\|\alpha\|=1, \|\alpha\|_0 \leq k} \mathbb{E}_n\left[(\alpha' x_i)^2\right] \simeq_p 1 + \sqrt{\frac{k \log p}{n}}$$

by Lemma 14. Thus, this design satisfies our conditions with $\phi(n/\log(n \vee p)) \simeq_p 1$. Moreover, as shown in [8], this design satisfies Candes and Tao's uniform uncertainty principle.

Example 2 (Correlated Normal Design). We consider the same setup as in Example 1, but instead we suppose that the covariates are correlated, namely $x_{-1} \sim N(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$ and $-1 < \rho < 1$ is fixed. This example satisfies conditions D.1, D.3. D.4, and D.2 if $\|\beta_0\|_0 \leq s = o(n/\log(n \vee p))$. The maximal k-sparse eigenvalues for $k \leq n$ satisfy

$$\phi(k) := \sup_{\|\alpha\|=1, \|\alpha\|_0 \leq k} \mathbb{E}_n\left[(\alpha' x_i)^2\right] \simeq_p \frac{1 + |\rho|}{1 - |\rho|}\left(1 + \sqrt{\frac{k \log p}{n}}\right)$$

by Lemmas 14. Thus, this design satisfies our conditions with $\phi(n/\log(n \vee p)) \simeq_p 1$. However, as mentioned in [26] this design violates Candes and Tao's uniform uncertainty principle, which requires $|\rho| \to 0$ at $\log p$ rate.

Finally, it is worth noting that our analysis in Sections 3 and 4, and in Appendix A allows the key parameters of the model, such as the bounds on the eigenvalues of the design matrix and on the density function, to change with the sample size. This will explicitly allow us to trace out the impact of these parameters on the large sample behavior of the penalized estimator. In particular, we will be able to immediately see how some basic changes in the primitive conditions stated above affect the large sample behavior of the penalized estimator.

2.5. *Overview of Main Results.* Here we discuss our results under simplest assumptions, consisting of conditions D.1-D.4 and condition (2.15) on the maximal $(n/\log(n \vee p))$-sparse eigenvalue. These simplest assumptions allow us to straightforwardly compare our results to those obtained in the literature, without getting into nuisance details. We state our results under more general conditions in the subsequent sections: in Section 3, we present various results on convergence rates and model selection; in Section 4, we analyze our choice of the penalization parameter.

In order to achieve the most rapid rate of convergence, we need to choose

$$\lambda = t\sqrt{n \log(n \vee p)} \tag{2.16}$$

with t growing as slowly as possible with n; for concreteness, let $t \propto \log\log n$.

Our first main result is that the ℓ_1-penalized quantile regression estimator converges at the rate:

$$\|\widehat{\beta}(u) - \beta(u)\| \lesssim_P \frac{\lambda\sqrt{s}}{n} = \sqrt{\frac{s}{n}} \cdot t \cdot \sqrt{\log(n \vee p)}, \tag{2.17}$$

provided that the number of non-zero components s satisfies

$$\sqrt{\frac{s}{n}} \cdot t \cdot \sqrt{\log(n \vee p)} \to 0. \tag{2.18}$$

We note that the total number of regressors p affects the rate of convergence (2.17) only through a logarithm in p. Hence if p is polynomial in n, the rate of convergence is $\sqrt{s/n} \cdot t \cdot \sqrt{\log(n \vee p)}$, which is very close to the oracle rate $\sqrt{s/n}$, obtainable when we know the minimal true model. Further, we note that our resulting restriction (2.18) on the dimension s of the minimal true model is very weak; when p is polynomial in n and $t \propto \log\log n$, s can be of almost the same order as n, namely $s = o(n/(t^2 \log n)))$.

Our second main result is that the dimension $\|\widehat{\beta}(u)\|_0$ of the model selected by the ℓ_1-penalized estimator is of the same stochastic order as the dimension s of the minimal true model, namely

$$\|\widehat{\beta}(u)\|_0 \lesssim_p s. \tag{2.19}$$

Further, if the parameter values of the minimal true model are well separated away from zero, namely

$$\min_{j \in \text{support}(\beta(u))} |\beta_j(u)| > \ell \cdot \sqrt{\frac{s}{n}} \cdot t \cdot \sqrt{\log(n \vee p)}, \tag{2.20}$$

for some diverging sequence ℓ of positive constants, then with probability converging to one, the model selected by the ℓ_1-penalized estimator correctly nests the true minimal model:

$$\text{support } (\beta(u)) \subseteq \text{support } (\widehat{\beta}(u)). \tag{2.21}$$

Moreover, we provide conditions under which a hard-thresholding selects the correct support.

Our third main result is that a two-step estimator, which applies standard quantile regression to the selected model, achieves a similar rate of convergence:

$$\sqrt{\frac{s}{n}}\sqrt{\log(n \vee p)} \to 0, \tag{2.22}$$

provided the true non-zero coefficients are well-separated from zero in the sense of equation (2.20).

Finally, our fourth main result is to propose (2.13), a data-driven choice of the regularization parameter λ which has a pivotal finite sample distribution conditional on the regressors, and to verify that (2.13) satisfies the theoretical restriction (2.16), supporting its use in practical estimation.

Our results for quantile regression parallel the results for least squares by Meinshausen and Yu [26] and by Candes and Tao [8]. Our results on the pivotal choice of the regularization parameter partly parallel the results by Candes and Tao [8], except that our choice is pivotal whereas Candes and Tao's choice relies upon the knowledge of the standard deviation of the regression disturbances. The existence of close parallels may seem surprising, since, in contrast to the least squares problem, our problem is highly non-linear and non-smooth. Nevertheless, there is an intuition presented below, suggesting that we can overcome these difficulties.

While our results for quantile regression parallel results for least squares, our proof strategy is substantially different, as it has to address non-linearities and non-smoothness. In order to explain the difference, let us recall, e.g., the proof strategy of Meinshausen and Yu [26]. They first analyze the problem with no disturbances, recognize sparseness of the solution for this zero noise problem, and then analyze a sequence of problems along the path interpolating the zero-noise problem and the full-noise problem. Along this sequence, they bound the increments in the number of non-zero components and in the rates of convergence. This approach does not seem to work for our problem, where the zero-noise problem does not seem to have either the required sparseness or the required smoothness. In sharp contrast, our approach directly focuses on the full-noise problem, and simultaneously bounds the number of non-zero components and convergence rates. Thus, our approach may be of independent interest for other M-estimation problems and even for the least squares problem.

Our analysis is perhaps closer in spirit to, but still quite different from, the important work of van der Geer [37] which derived finite sample bounds on the empirical risk of ℓ_1-penalized estimators in generalized linear models (but did not investigate quantile regression models). The major difference between our proof and van der Geer [37]'s proof strategies is that we analyze the sparseness of the solution to the penalized problem and then further exploit sparseness to control empirical errors in the sample criterion function. As a result, we derive not only the results on model selection and on sparseness of solutions, which are of a prime interest, but also the results on the consistency and rates of convergence under weak conditions on the number of non-zero components s. As mentioned above, our approach allows s to be of almost the same order as the sample size n, and delivers convergence rates that are close to $\sqrt{s/n}$. In contrast, van der Geer's [37] approach requires s to be much smaller than n, namely $s^2/n \to 0$, and thus does not deliver consistency or rates of convergence when $s^2/n \to \infty$.

In our proofs we critically rely on two key quantities: the number of non-zero components $m = \|\widehat{\beta}(u)\|_0$ of the solution $\widehat{\beta}(u)$ and the empirical error in the sample criterion, $\widehat{Q}_u(\beta) - Q_u(\beta)$ (with β ranging over all m-dimensional submodels of the large p-dimensional model). In particular, we make use of the following relations:

(1) lower m implies smaller empirical error, and

(2) smaller empirical error can imply lower m.

Starting with a structural initial upper bound on m (see condition E.3 below) we can use the two relations to solve for sharp bounds on m and the empirical error, given which we then can solve for convergence rates.

Let us comment on the intuition behind relations (1) and (2). Relation (1) follows from an application of the usual entropy-based maximal inequalities, upon realizing that the entropy of all m dimensional models grows at the rate $m \log p$. In particular, the lower the m, the closer the sample criterion function $\widehat{Q}_u$ to a locally quadratic function, uniformly across all m-dimensional submodels. Relation (2) follows from the use of ℓ_1-penalty, which tends to favor lower-dimensional solutions when $\widehat{Q}_u$ is close to being quadratic. Figure 1 provides a visual illustration of this, using a two-dimensional example with a one-dimensional true minimal submodel; in the example, the true parameter value $(\beta_1(u), \beta_2(u))$ is $(1, 0)$. Figure 1 plots a diamond, centered at the origin, representing a contour set of the ℓ_1-penalty function and a pearl, representing a contour set of the criterion function $\widehat{Q}_u$. By the dual interpretation (2.9) of our estimation problem, the penalized estimator looks for a minimal diamond, subject to the diamond having a non-empty intersection with a fixed pearl. The set of optimal solutions is then given by the intersection of the minimal diamond with the pearl. Smaller empirical errors shape the pearl into an ellipse and center it closer to the true parameter value of $(1, 0)$ (left panel of Figure 1). Larger empirical errors shape the pearl like a non-ellipse and can center it far away from the true parameter value (right panel of Figure 1). Therefore, smaller empirical errors tend to cause sparse optimal solutions, correctly setting $\widehat{\beta}_2(u) = 0$; larger empirical errors tend to cause non-sparse optimal solutions, incorrectly setting $\widehat{\beta}_2(u) \neq 0$.

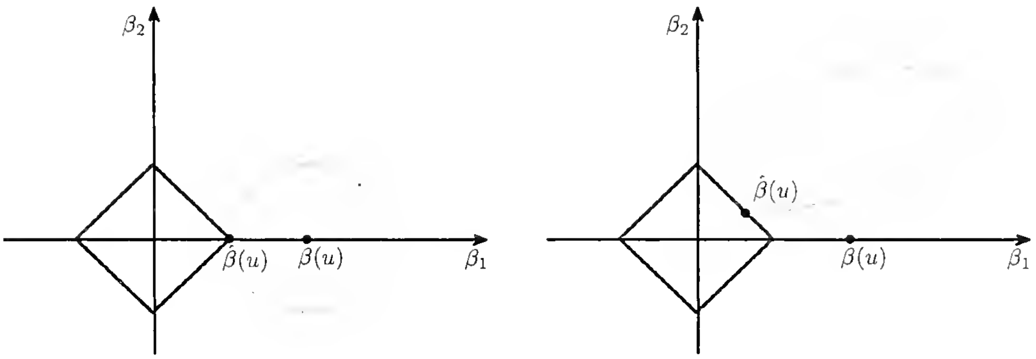

FIG 1. *These figures provide a geometric illustration for the discussion given in the text concerning why ℓ_1-penalized estimation may be (left panel) or may not be (right panel) successful at selecting the minimal true model.*

3. Analysis and Main Results Under High-Level Conditions. In this section we prove the main results under general conditions that encompass the simple conditions

D.1-D.4 as a special case.

3.1. *The Five Basic Conditions.* We will work with the following five basic conditions E.1-E.5 which are the essential ingredients needed for our asymptotic approximations. In Appendix A, we verify that conditions E.1-E.5 hold under simple sufficient conditions D.1-D.4 stated in Section 2, and we also show that E.1-E.5 arise much more generally. In particular, in Appendix A we characterize key constants appearing in E.1-E.5 in terms of the parameters of the model.

E.1. True Model Sparseness. The true parameter value $\beta(u)$ has at most $s < n/\log(n \vee p)$ non-zero components, namely

$$\|\beta(u)\|_0 = s < n/\log(n \vee p). \tag{3.1}$$

E.2. Identification in Population. In the population, the true parameter value $\beta(u)$ is the unique solution to the quantile objective function. Moreover, the following minorization condition holds,

$$Q_u(\beta) - Q_u(\beta(u)) \gtrsim q\left(\|\beta - \beta(u)\|^2 \wedge g(\|\beta - \beta(u)\|)\right), \tag{3.2}$$

uniformly in $\beta \in \mathrm{IR}^p$, where $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a fixed convex function with $g'(0) > 0$, and q is a sequence of positive numbers that characterizes the strength of identification in the population.

E.3. Empirical Pre-Sparseness. The number $m = \|\widehat{\beta}(u)\|_0$ of non-zero components of $\widehat{\beta}(u)$ of the solution to the penalized quantile regression problem (2.4) obeys the inequality

$$m \leq n \wedge p \wedge \frac{n^2\phi(m)}{\lambda^2}, \tag{3.3}$$

where $\phi(m)$ is the maximal m-sparse eigenvalue.

E.4. Empirical Sparseness. For $r = \|\widehat{\beta}(u) - \beta(u)\|$, $m = \|\widehat{\beta}(u)\|_0$ obeys the following stochastic inequality

$$\sqrt{m} \lesssim_p \mu\frac{n}{\lambda}(r \wedge 1) + \sqrt{m}\frac{\sqrt{n\log(n \vee p)\phi(m)}}{\lambda}, \tag{3.4}$$

where $\mu \geq q$ is a sequence of positive constants. The sequence of constants μ is determined by the population analog of the empirical sparse eigenvalue $\phi(m_0)$ (cf. Appendix A).

E.5. Sparse Control of Empirical Error. The empirical error that describes the deviation of the empirical criterion from the population criterion satisfies the following stochastic inequality

$$(3.5)\qquad \left|\widehat{Q}_u(\beta) - Q_u(\beta) - \left(\widehat{Q}_u(\beta(u)) - Q_u(\beta(u))\right)\right| \lesssim_P r\sqrt{\frac{(m+s)\log(n \vee p)\phi(m+s)}{n}},$$

uniformly over $\{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq m \wedge n \wedge p,\ \|\beta - \beta(u)\| \leq r\}$, uniformly over $m \leq n, r \geq 0$.

Let us briefly comment on each of the conditions. As stated earlier, condition E.1 is a basic modeling assumption, and condition E.2 is an identification assumption, required to hold in population. Conditions E.3 and E.4 arise from two characterizations of sparseness of the solution to the optimization problem (2.4) defining the estimator. Condition E.3 arises from simple bounds applied to the first characterization. Condition E.4 arises from maximal inequalities applied to the second characterization. Condition E.5 arises from maximal inequalities applied to the empirical criterion function. To derive conditions E.4 and E.5, we crucially exploit the fact that the entropy of all m-dimensional submodels of the p-dimensional model is of order $m \log p$, which depends on p only logarithmically. Finally, we note that Conditions E.1-E.5 easily hold under primitive assumptions D.1-D.4, in particular $\mu \simeq q \simeq 1$, but we also permit them to hold more generally. We refer the reader to Section 5 for verification and further analysis of these conditions.

Theorem 1 combines conditions E.1-E.5 to establish bounds on the rate of convergence and sparseness of the estimator (2.4).

THEOREM 1. *Assume that conditions E.1-E.5 hold. Let $t \to_P \infty$ be a sequence of positive numbers, possibly data-dependent, define*

$$(3.6)\qquad m_0 = p \wedge \left(\frac{n}{\log(n \vee p)}\frac{q^2}{\mu^2}\right), \text{ and set } \lambda = t\sqrt{n\log(n \vee p)\phi(m_0+s)}\frac{\mu}{q}.$$

Then we have that

$$(3.7)\qquad \|\widehat{\beta}(u) - \beta(u)\| \lesssim_P \frac{\lambda\sqrt{s}}{qn} = t\sqrt{\frac{s\log(n \vee p)\phi(m_0+s)}{n}}\frac{\mu}{q^2},$$

provided that $\lambda\sqrt{s}/(qn) \to_P 0$, and

$$(3.8)\qquad \|\widehat{\beta}(u)\|_0 \lesssim_P \left(\frac{\mu}{q}\right)^2 s.$$

This is the main result of the paper that derives the rate of convergence of the ℓ_1-penalized quantile regression estimator and a stochastic bound on the dimension of the selected model. Our results parallel the results of Meinshausen and Yu [26] obtained for the ℓ_1-penalized mean regression. We refer the reader to Section 2 for a detailed discussion of this and other main results of this section under simplified conditions. Here we only note that the rate of convergence generally depends on the number of significant regressors s, the logarithm of the number of regressors p, the strength of identification q, the empirical sparse eigenvalue $\phi(m_0)$, and the constant μ determined by the population sparse eigenvalue. The bound on the dimension also depends on the sequence of constants s, q, and μ.

It is also helpful to state the main result separately under the simple set of conditions D.1-D.4, where $q \simeq \mu \simeq 1$.

Corollary 1 (A Leading Case). *Conditions D.1-D.4 imply conditions E.1-E.5 with $q \simeq \mu \simeq 1$. Therefore, under D.1-D.4, $m_0 = p \wedge (n/\log(n \vee p))$, so setting*

$$\lambda = t\sqrt{n\log(n \vee p)\phi(m_0)} \quad \textit{and if} \quad t\sqrt{\frac{s\log(n \vee p)\phi(m_0)}{n}} \to 0$$

we have that

$$\|\widehat{\beta}(u) - \beta(u)\| \lesssim_p t\sqrt{\frac{s\log(n \vee p)\phi(m_0)}{n}},$$

and

$$\|\widehat{\beta}(u)\|_0 \lesssim_p s.$$

If in addition $\phi(m_0) \lesssim_p 1$, then we obtain the rate result listed in equation (2.17).

This corollary follows from lemmas stated in Appendix A, where we verify that conditions D.1-D.4 imply conditions E.1-E.5. Moreover, we use the fact that $\phi(m_0 + s) \leq \phi(2m_0)$ if $s\log(n \vee p) < n$ for $m_0 = p \wedge (n/\log(n \vee p))$, and that $\phi(2m_0) \leq 2\phi(m_0)$ by Lemma 11.

It is useful to revisit our concrete examples.

Example 3 (Isotropic Normal Design, continued). In the isotropic normal design considered earlier, recall that we have that $\phi(k) \lesssim_p 1 + \sqrt{(k/n)\log p}$. If $\lambda/\sqrt{n\log(n \vee p)} \to \infty$, by Theorem 1 we have $m_0 \leq n/\log(n \vee p)$, and, since we assume $s \leq n/\log(n \vee p)$, by Lemma 11 we have $\phi(m_0 + s) \lesssim_p 1$. Also, we verify in Appendix A that $q \simeq \mu \simeq 1$. Thus, the rate result listed in equation (2.17) applies to this example.

Example 4 (Correlated Normal Design, continued). In the correlated normal design considered earlier, we have that $\phi(k) \lesssim_p \frac{1+|\rho|}{1-|\rho|}(1 + \sqrt{(k/n)\log p})$. If $\lambda/\sqrt{n\log(n \vee p)} \to \infty$, by Theorem 1 we have $m_0 \leq n/\log(n \vee p)$ and, since we assume $s \leq n/\log(n \vee p)$, by Lemma

11 we have $\phi(m_0+s) \lesssim_P \frac{1+|\rho|}{1-|\rho|} \lesssim_P 1$. Also, we verify in Appendix A, that $q \simeq \mu \simeq 1$. Thus, the rate result listed in equation (2.17) applies to this example too.

PROOF. (Theorem 1) Let

$$r := \|\widehat{\beta}(u) - \beta(u)\| \text{ and } m := \|\widehat{\beta}(u)\|_0.$$

The proof successively refines upper bounds on m and r. We divide the proof in four steps. The first step provides an initial bound on m, the second step obtains preliminary inequalities, the third step verifies consistency, and the fourth step establishes the rate result.

STEP 1. We start by proving that $m \leq m_0$ if $t \geq \sqrt{2}$. Since $t \to_P \infty$, $m \leq m_0$ will occur with probability converging to one. By condition E.3 we have

$$m \leq \bar{m} = \max\left\{ m : m \leq n \wedge p \wedge \frac{n^2\phi(m)}{\lambda^2} \right\}.$$

If $m_0 = p$ we have directly that $\bar{m} \leq m_0$. Next consider the case $m_0 = \left(\frac{n}{\log(n \vee p)} \frac{q^2}{\mu^2} \right)$.

Suppose that $\bar{m} > m_0$ when $t \geq \sqrt{2}$. Therefore we have $\bar{m} = m_0 \ell$ for some $\ell > 1$ (since $\bar{m} \leq n \wedge p$ is finite). By definition $\bar{m}$ satisfies the inequality

$$\bar{m} \leq n^2 \frac{\phi(\bar{m})}{\lambda^2}. \tag{3.9}$$

Since $\phi(m_0) \leq \phi(m_0+s)$ we have $\lambda \geq t\sqrt{n \log(n \vee p)\phi(m_0)}(\mu/q)$. Inserting this bound on λ, the value of m_0, and $\bar{m} = m_0\ell$ in (3.9), and then using Lemma 11 and $t \geq \sqrt{2}$ we obtain

$$\bar{m} = m_0\ell \leq \frac{n^2}{t^2 n \log(n \vee p)} \frac{\phi(m_0\ell)}{\phi(m_0)} \frac{q^2}{\mu^2} < \frac{n}{t^2 \log(n \vee p)} 2\ell \frac{q^2}{\mu^2} = \frac{2}{t^2} m_0 \ell \leq m_0 \ell,$$

which is a contradiction.

STEP 2. In this step we obtain some preliminary inequalities.

By Condition E.1, the support of $\beta(u)$

$$T_u := \text{ support}(\beta(u)) := \{j \in \{1, \ldots, p\} : \ |\beta_j(u)| > 0\}$$

has exactly s elements, that is, $|T_u| = s$. Let $\widehat{\beta}_{T_u}(u)$ denote a vector whose T_u components agree with T_u components of $\widehat{\beta}(u)$, and whose remaining components are equal to zero.

By definition of $\widehat{\beta}(u)$ and since $\|\widehat{\beta}_{T_u}(u)\|_1 \leq \|\widehat{\beta}(u)\|_1$ we have that

$$\widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u)) \leq \frac{\lambda}{n}(\|\beta(u)\|_1 - \|\widehat{\beta}(u)\|_1) \leq \frac{\lambda}{n}(\|\beta(u)\|_1 - \|\widehat{\beta}_{T_u}(u)\|_1).$$

Using that

$$|\|\beta(u)\|_1 - \|\widehat{\beta}_{T_u}(u)\|_1| \leq \|\beta(u) - \widehat{\beta}_{T_u}(u)\|_1 \leq \sqrt{|T_u|}\|\widehat{\beta}_{T_u}(u) - \beta(u)\| \leq \sqrt{s}r$$

we obtain that

$$\widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u)) \leq \frac{\lambda}{n}\sqrt{s}r.$$

Applying condition E.5 to control the difference between the sample and population criterion functions, we further get that

$$Q_u(\widehat{\beta}(u)) - Q_u(\beta(u)) \lesssim_p \frac{\lambda}{n}\sqrt{s}r + r\sqrt{\frac{(m+s)\log(n \vee p)\phi(m+s)}{n}}.$$

Invoking the identification condition E.2 and the definition of r, we obtain

$$q(r^2 \wedge g(r)) \lesssim_p \frac{\lambda}{n}\sqrt{s}r + r\sqrt{\frac{(m+s)\log(n \vee p)\phi(m+s)}{n}}. \tag{3.10}$$

Step 3. In this step we show consistency, namely $r = o_p(1)$. By Step 1 we have $m \leq m_0$ with probability converging to one.

The construction (3.6) of λ, $t \to_p \infty$, and the condition $\lambda\sqrt{s}/(qn) \to_p 0$ assumed in the theorem imply

$$(i)\frac{\lambda\sqrt{s}}{n} = o_p(q), \ (ii)\sqrt{\frac{s\log(n \vee p)\phi(m_0+s)}{n}}\frac{\mu}{q} = o_p(q), \ \ (iii)\mu\frac{\sqrt{n\log(n \vee p)\phi(m_0+s)}}{\lambda} = o_p(q).$$

Condition (iii), $\mu \geq q$, and empirical sparseness condition E.4, stated in equation (3.4), imply that

$$\sqrt{m} \lesssim_p \mu(r \wedge 1)n/\lambda + \sqrt{m}o_p(1), \tag{3.11}$$

which implies the following second bound on m:

$$\sqrt{m} \lesssim_p \mu n/\lambda. \tag{3.12}$$

Using (3.12) and $m \geq s$ in equation (3.10) gives

$$1\{m > s\}q\left(r^2 \wedge g(r)\right) \lesssim_p r\frac{\lambda}{n}\sqrt{s} + r\frac{\sqrt{n\log(n \vee p)\phi(m_0+s)}\mu}{\lambda} = ro_p(q) \tag{3.13}$$

where the last equality follows by conditions (i) and (iii). On the other hand, using (3.12) and $m \leq s$ in equation (3.10) gives

$$1\{m \leq s\} q\left(r^2 \wedge g(r)\right) \lesssim_p r\frac{\lambda}{n}\sqrt{s} + r\sqrt{\frac{s\log(n \vee p)\phi(m_0+s)}{n}} = r o_p(q) \tag{3.14}$$

where the last equality follows by conditions (i) and (ii) and $\mu \geq q$. Conclude from (3.13) and (3.14) that

$$q\left(r^2 \wedge g(r)\right) = r o_p(q). \tag{3.15}$$

Next we show that (3.15) implies $r = o_p(1)$. Dividing both sides of (3.15) by q and by r we have $1\{r > 0\}[r \wedge (g(r)/r)] \lesssim_p 1\{r > 0\}o_p(1)$. By condition E.2, g is a fixed convex function with $g'(0) > 0$, so that $g(r) \geq g'(0)r$. Thus, $1\{r > 0\}[r \wedge g'(0)] = 1\{r > 0\}o_p(1)$, that is, $r = o_p(1)$.

STEP 4. This step derives the rate of convergence.

Using that $r = o_p(1)$ we improve the bound (3.11) on m to the following third bound:

$$\sqrt{m} \lesssim_p \frac{r\mu n}{\lambda}. \tag{3.16}$$

Plugging (3.16) into (3.10) and using the relation $r^2 = o_p(g(r))$ under $r = o_p(1)$, gives us

$$qr^2 \lesssim_p r\frac{\lambda\sqrt{s}}{n} + o_p(q)r^2 \text{ or equivalently } r \lesssim_p \frac{\lambda\sqrt{s}}{qn}. \tag{3.17}$$

Finally, inserting (3.17) into (3.16), we obtain $\sqrt{m} \lesssim_p \sqrt{s}(\mu/q)$, which verifies the final bound (3.8) on m. □

3.2. *Model Selection Properties.* Next we turn to the model selection properties of the estimator.

THEOREM 2. *If conditions of Theorem 1 hold, and if the non-zero components of $\beta(u)$ are separated away from zero, namely*

$$\min_{j \in support(\beta(u))} |\beta_j(u)| > \ell t\sqrt{\frac{s\log(n \vee p)\phi(m_0+s)}{n}}\frac{\mu}{q^2}, \tag{3.18}$$

for some diverging sequence ℓ of positive constants, $\ell \to \infty$, then with probability approaching one

$$support\ (\beta(u)) \subseteq support\ (\widehat{\beta}(u)). \tag{3.19}$$

Moreover, the hard-thresholded estimator $\bar{\beta}(u)$, defined by

$$\bar{\beta}_j(u) = \widehat{\beta}_j(u) 1 \left\{ |\widehat{\beta}_j(u)| > \ell' t \sqrt{\frac{s \log(n \vee p) \phi(m_0 + s)}{n} \frac{\mu}{q^2}} \right\}$$

where $\ell' \to \infty$ and $\ell'/\ell \to 0$, satisfies with probability converging to one,

$$\textit{support } (\bar{\beta}(u)) = \textit{support } (\beta(u)).$$

Theorem 2 derives some model selection properties of the ℓ_1−penalized quantile regression. These results parallel analogous results obtained by Meinshausen and Yu [26] for the ℓ_1-penalized mean regression. The first result says that in order for the support of the estimator to include the support of the true model, non-zero coefficients need to be well-separated from zero, which is a stronger condition than what we required for consistency. The inclusion of the true support is in general one-sided; the support of the estimator can include some unnecessary components having the true coefficients equal zero. The second result describes the performance of the ℓ_1-penalized estimator with an additional hard thresholding, which does eliminate inclusions of such unnecessary components. However, the value of the right threshold explicitly depends on the parameter values characterizing the separation of non-zero coefficients from zero.

Proof. (Theorem 2) The result on inclusion of the support stated in equation (3.19) follows from the separation assumption (3.18) and the inequality $\|\widehat{\beta}(u) - \beta(u)\|_\infty \leq \|\widehat{\beta}(u) - \beta(u)\|$. Indeed, by Theorem 1 we have with probability going to one,

$$\|\widehat{\beta}(u) - \beta(u)\|_\infty \leq \|\widehat{\beta}(u) - \beta(u)\| < \min_{j \in \text{support}(\beta(u))} |\beta_j(u)|. \tag{3.20}$$

The last inequality follows from the rate result of Theorem 1 and from the separation assumption (3.18). Next, the converse of the inclusion event (3.19) implies that $\|\widehat{\beta}(u) - \beta(u)\|_\infty \geq \min_{j \in \text{support}(\beta(u))} |\beta_j(u)|$. Since the latter can occur only with probability approaching zero, we conclude that the event (3.19) occurs with probability converging to one.

Consider the hard-thresholded estimator next. Let $r_n = t\sqrt{(s/n)\log(n \vee p)\phi(m_0 + s)}\mu/q^2$. To establish the inclusion note that by Theorem 1 and the separation assumption (3.18)

$$\min_{j \in \text{support } (\beta(u))} |\widehat{\beta}_j(u)| \geq \min_{j \in \text{support } (\beta(u))} \{|\beta_j(u)| - |\beta_j(u) - \widehat{\beta}_j(u)|\} \gtrsim_p \ell r_n - r_n$$

so that $\min_{j \in \text{support } (\beta(u))} |\widehat{\beta}_j(u)| > \ell' r_n$ with probability going to one by $\ell' \to \infty$ and $\ell'/\ell \to 0$. Therefore, support $(\beta(u)) \subseteq$ support $(\bar{\beta}(u))$ with probability going to one. To establish the opposite inclusion, consider the quantity

$$e_n = \max_{j \notin \text{support } (\beta(u))} |\widehat{\beta}_j(u)|.$$

By Theorem 1 $e_n \lesssim_p r_n$ so that $e_n < \ell' r_n$ with probability going to one by $\ell' \to \infty$. Since by the hard-threshold rule all components smaller than $\ell' r_n$ are excluded from the support of $\bar{\beta}(u)$, we have that support $(\bar{\beta}(u)) \subseteq$ support $(\beta(u))$ with probability going to one. □

3.3. *Two-step estimator.* Next we consider the following two-step estimator that applies the ordinary quantile regression to the selected model. Let $\widehat{T}$ be a model, that is, a subset of $\{1, \ldots, p\}$, selected by a data-dependent procedure. We define the two-step estimator $\widehat{\beta}^{\widehat{T}}(u)$ as a solution of the following optimization problem:

$$\widehat{\beta}^{\widehat{T}}(u) \in \arg \min_{\beta \in \mathbb{R}^p : \beta_j = 0, j \notin \widehat{T}} \widehat{Q}_u(\beta). \tag{3.21}$$

In this problem we constrain the components of the parameter vector β that were not selected to be zero; or, equivalently, we remove the regressors that were not selected from further estimation. Moreover, we no longer use ℓ_1-penalization.

THEOREM 3. *Suppose that conditions E.1, E.2, and E.5 hold. Let $\widehat{T}$ be any selected model that contains the true model T_u with probability converging to one, and whose dimension $|\widehat{T}|$ is of stochastic order s, then*

$$\left\| \widehat{\beta}^{\widehat{T}}(u) - \beta(u) \right\| \lesssim_p \sqrt{\frac{s \log(n \vee p) \phi(s)}{n}} \frac{1}{q},$$

provided the right side converges to zero in probability.

Under conditions of the theorem see that the rate of convergence of the two-step estimator is generally faster than the rate of the one-step penalized estimator, unless $\phi(n) \simeq_p \phi(s)$, in which case the rate is the same. It is also helpful to note that when $q \simeq 1$ and $\phi(s) \lesssim_p 1$,

$$\|\widehat{\beta}^{\widehat{T}}(u) - \beta(u)\| \lesssim_p \sqrt{\frac{s}{n} \log(n \vee p)}.$$

PROOF. (Theorem 3). Let $r = \|\widehat{\beta}^{\widehat{T}}(u) - \beta(u)\|$. By definition of $\widehat{\beta}^{\widehat{T}}(u)$ and by $T_u \subseteq \widehat{T}$ with probability approaching one, we have that with probability approaching one

$$\widehat{Q}_u(\widehat{\beta}^{\widehat{T}}(u)) - \widehat{Q}_u(\beta(u)) \leq 0.$$

First note that since $|\widehat{T}| \lesssim_p s$, by Lemma 11 we have that $\phi(|\widehat{T}|+s) \lesssim_p \phi(s)$. Applying condition E.5 to control the empirical error in the objective function, we get that

$$Q_u(\widehat{\beta}^{\widehat{T}}(u)) - Q_u(\beta(u)) \quad \lesssim_p \quad r\sqrt{\frac{s\log(n\vee p)\phi(|\widehat{T}|+s)}{n}} \lesssim_p r\sqrt{\frac{s\log(n\vee p)\phi(s)}{n}}.$$

Invoking the identification condition E.2 we obtain that

$$q(r^2 \wedge g(r)) \lesssim_p r\sqrt{\frac{s\log(n\vee p)\phi(s)}{n}}. \tag{3.22}$$

Since we assumed that $\sqrt{s\log(n\vee p)\phi(s)/n} = o_p(q)$, we conclude that $q(r^2 \wedge g(r)) \lesssim_p r o_p(q)$. As in the proof of Theorem 1, this implies that $r = o_p(1)$, and that $r^2 = o_p(g(r))$. Therefore we can refine the bound (3.22) to

$$qr^2 \lesssim_p r\sqrt{\frac{s\log(n\vee p)\phi(s)}{n}} \quad \text{or} \quad r \lesssim_p \sqrt{\frac{s\log(n\vee p)\phi(s)}{n}}\frac{1}{q},$$

proving the result. □

4. Analysis of the Pivotal Choice of the Penalization Parameter.

In this section we show that under some conditions the pivotal choice for the penalization parameter λ proposed in Section 2.3 satisfies the theoretical requirements needed to achieve the rates of convergence stated in Theorem 1.

Recall that the true rank scores can be represented almost surely as

$$a_i^*(u) = (u - 1\{u_i \le u\}), \quad \text{for } i = 1, \dots, n,$$

where $u_1, \dots, u_n$ are i.i.d. uniform $(0,1)$ random variables, independently distributed from the regressors, $x_1, \dots, x_n$. Thus, we have

$$\Lambda_n = n \left\| \mathbb{E}_n \left[x_i(u - 1\{u_i \le u\}) \right] \right\|_\infty, \tag{4.1}$$

which has a known distribution conditional on $X = (x_1, \dots, x_n)$.

THEOREM 4. *Let the regularization parameter $\lambda(X)$ be defined as*

$$\lambda(X) = \inf\{\lambda : P(\Lambda_n \le \lambda | X) \ge 1 - \alpha_n\}, \quad \alpha_n = \left(\frac{1}{n\vee p}\right)^{t^2}, \tag{4.2}$$

for some sequence $t \to \infty$. Assume that there exists a sequence $c_{n,p}$ such that uniformly in $j = 1, \dots, p$

$$\max_{i=1,\dots,n} |x_{ij}| \leq c_{n,p} \left(\sum_{i=1}^{n} x_{ij}^2 \right)^{1/2} \quad \textit{and} \quad c_{n,p} \cdot t \cdot \sqrt{\log(n \vee p)} \to 0. \tag{4.3}$$

Moreover, assume $q \simeq \mu$, $\phi(1) \simeq_p \phi(n/\log(n \vee p))$, and that $t\sqrt{s \log(n \vee p)\phi(1)/n}/q \to_p 0$. Then $\lambda = \lambda(X)$ satisfies the assumptions on the regularization parameter assumed in Theorem 1, namely there exists a sequence $\tilde{t} \to_p \infty$ such that

$$\lambda = \tilde{t}\sqrt{n \log(n \vee p)\phi(m_0 + s)}\frac{\mu}{q} \quad \textit{and} \quad \frac{\lambda\sqrt{s}}{qn} \to_p 0 \tag{4.4}$$

where $m_0 = p \wedge \left(\dfrac{n}{\log(n \vee p)} \dfrac{q^2}{\mu^2} \right)$, and $\tilde{t} \simeq_p t$.

Proof. (Theorem 4) We will use the following inequalities of Stout [34], Theorem 5.2.2: Let $\{X_i, i \geq 1\}$ denote a sequence of independent random variables with zero mean and finite variances, and let $S_n = \sum_{i=1}^n X_i$ and $s_n^2 = \sum_{i=1}^n \mathrm{E}\left[X_i^2\right]$ for all $n \geq 1$. Let $|X_i| \leq cs_n$ almost surely for each $1 \leq i \leq n$ and $n \geq 1$. Suppose $\varepsilon > 0$ and $\gamma > 0$. Then for each $n \geq 1$, the inequality $\varepsilon c \leq 1$ implies that

$$P\left(S_n/s_n > \varepsilon\right) \leq \exp\left(-\left(\varepsilon^2/2\right)(1 - \varepsilon c/2)\right), \tag{4.5}$$

and there exist constants $\varepsilon(\gamma)$ and $\pi(\gamma)$ such that if $\varepsilon \geq \varepsilon(\gamma)$ and $\varepsilon c \leq \pi(\gamma)$, then

$$P\left(S_n/s_n > \varepsilon\right) \geq \exp\left(-\left(\varepsilon^2/2\right)(1 + \gamma)\right). \tag{4.6}$$

We need to establish upper and lower bounds on the value of λ. We first establish an upper bound. Let $v_j^2 = \sum_{i=1}^n x_{ij}^2$ and note that $\phi(1) = \sup_{j \leq p} v_j^2/n$. Next observe that $\mathrm{Var}\left(\sum_{i=1}^n x_{ij} a_i^*(u) | X\right) = u(1-u)v_j^2$. Note that by (4.3) we have $\sup_{1 \leq i \leq n} |x_{ij} a_i^*(u)| \leq c_{n,p} v_j / \sqrt{u(1-u)}$, $j = 1, \dots, p$. Moreover, for n large enough, condition (4.3) also implies that

$$2c_{n,p}(t+1)\sqrt{\log(n \vee p)}/\sqrt{u(1-u)} < 1/2. \tag{4.7}$$

Under (4.7), we can apply (4.5) with $\varepsilon = 2(t+1)\sqrt{\log(n \vee p)}$, and $c = c_{n,p}/\sqrt{u(1-u)}$ to obtain that for every $j = 1, \dots, p$

$$P\left(\frac{|\sum_{i=1}^n x_{ij} a_i^*(u)|}{\sqrt{u(1-u)}v_j} > \varepsilon | X \right) \leq \exp\left(-\frac{\varepsilon^2}{2}\left(1 - \frac{c_{n,p}\varepsilon}{2\sqrt{u(1-u)}}\right)\right) < \exp\left(-(t^2+1)\log(n \vee p)\right). \tag{4.8}$$

Therefore, since $\sqrt{n\phi(1)} \geq v_j$ we have

$$P\left(\frac{|\sum_{i=1}^n x_{ij}a_i^*(u)|}{\sqrt{u(1-u)n\phi(1)}} > \varepsilon|X\right) < \exp\left(-(t^2+1)\log(n \vee p)\right) = \frac{1}{(n\vee p)}\left(\frac{1}{(n\vee p)}\right)^{t^2}. \tag{4.9}$$

Next note that using (4.9) we have

$$\begin{aligned} P\left(\Lambda_n > \sqrt{u(1-u)n\phi(1)}\varepsilon|X\right) &\leq \sum_{j=1}^p P\left(\left|\sum_{i=1}^n x_{ij}a_i^*(u)\right| > \sqrt{u(1-u)n\phi(1)}\varepsilon|X\right) \\ &\leq p\max_{j\leq p} P\left(\left|\sum_{i=1}^n x_{ij}a_i^*(u)\right| > \sqrt{u(1-u)n\phi(1)}\varepsilon|X\right) \\ &< \left(\tfrac{1}{(n\vee p)}\right)^{t^2}. \end{aligned} \tag{4.10}$$

Since $P(\Lambda_n > \lambda|X)$ is decreasing in λ, we conclude that

$$\lambda \leq \sqrt{u(1-u)n\phi(1)}\varepsilon \lesssim 2(t+1)\sqrt{n\log(n\vee p)\phi(1)}. \tag{4.11}$$

Next we turn to establishing the lower bound. Let $j_n \in \{1,\dots,p\}$ denote an index such that $v_{j_n} = \sqrt{n\phi(1)}$. By definition of Λ_n we have

$$\left(\frac{1}{(n\vee p)}\right)^{t^2} \geq P\left(\max_{j\leq p}\left|\sum_{i=1}^n x_{ij}a_i^*(u)\right| > \lambda|X\right) \geq P\left(\left|\sum_{i=1}^n x_{ij_n}a_i^*(u)\right| > \lambda|X\right).$$

Fix $\gamma > 0$ (which implicitly fix $\varepsilon(\gamma)$ and $\pi(\gamma)$), and set $\varepsilon = t\sqrt{2\log(n\vee p)/(1+\gamma)}$, $c = c_{n,p}/\sqrt{u(1-u)}$. Since ε diverges, and, by (4.3) we have $\varepsilon c = o(1)$, for n large enough we have $\varepsilon > \varepsilon(\gamma)$ and $\varepsilon c < \pi(\gamma)$. Therefore we can apply (4.6) to obtain

$$\begin{aligned} P\left(\frac{|\sum_{i=1}^n x_{ij_n}a_i^*(u)|}{\sqrt{u(1-u)n\phi(1)}} > \varepsilon|X\right) &\geq \exp\left(-(\varepsilon^2/2)(1+\gamma)\right) \\ &\geq \exp\left(-t^2\log(n\vee p)\right) = \left(\tfrac{1}{(n\vee p)}\right)^{t^2}. \end{aligned}$$

Since $P(\Lambda_n > \lambda|X)$ is decreasing in λ, it follows that

$$\lambda \geq \varepsilon\sqrt{u(1-u)n\phi(1)} = t\sqrt{2u(1-u)n\log(n\vee p)\phi(1)/(1+\gamma)}. \tag{4.12}$$

Thus, taking in account that $\mu \simeq q$, we have established $\lambda \simeq_p t\sqrt{n\log(n\vee p)\phi(m_0+s)}\frac{\mu}{q}$. In order to verify (4.4) define $\tilde{t} = \lambda/[\sqrt{n\log(n\vee p)\phi(m_0+s)}(\mu/q)]$. By construction we have that $\tilde{t} \simeq_p t \to \infty$. Thus, the first result of (4.4) follows, and the second result of (4.4) follows from the assumptions that $t\sqrt{s\log(n\vee p)\phi(1)/n}/q \to_p 0$. □

For concreteness, we now verify the conditions of Theorem 4 in our examples.

Example 5 (Isotropic Normal Design, continued). Let $x_{\cdot j}$ denote the n-vector associated with the jth covariate, where $x_{\cdot 1}$ is a column of ones representing the intercept. Next we use standard Gaussian concentration bounds, see [23] Section 3. For any value $K > 1$ we have

$$P(|x_{ij}| > K) \leq \exp(-K^2/2). \tag{4.13}$$

In turn this implies that $\max_{1\leq i\leq n, 1\leq j\leq p} |x_{ij}| \lesssim_P \sqrt{\log(n \vee p)}$. Moreover, the vectors $x_{\cdot j}$ are such that

$$P\left(| \, \|x_{\cdot j}\| - \mathrm{E}\left[\|x_{\cdot j}\|\right] \, | > K \right) \leq 2\exp\left(-2K^2/\pi^2\right) \text{ and } \mathrm{E}\left[\|x_{\cdot j}\|\right] \simeq \sqrt{n}, j = 1, \ldots, p. \tag{4.14}$$

Combining these bounds we obtain $\min_{j=1,\ldots,p} \sqrt{\sum_{i=1}^n x_{ij}^2} \gtrsim_P \sqrt{n} - \sqrt{\log p}$. Therefore, conditions (4.3) hold with $c_{n,p} \simeq_P \sqrt{\frac{\log(n\vee p)}{n}}$ and $t^2 \log^2(n \vee p) = o(n)$. On the other hand, we have $\phi(1) \geq 1$ and $\phi(m_0 + s) \lesssim_P 1 + \sqrt{(m_0/n)\log p} + \sqrt{(s/n)\log p} \lesssim 1$ by Lemma 14 and the definition of m_0. Thus, Theorem 4 requires $t^2 s \log(n \vee p) = o(n)$. We also verify that $q \simeq \mu \simeq 1$ in the next section.

Example 6 (Correlated Normal Design, continued). We analyze the correlated design similarly using comparison theorems for Gaussian random variables, Corollary 3.12 of Ledoux and Talagrand [23]. The upper bound for the case $\rho > 0$ follows from the result that for $K > 1$

$$P\left(\max_{1\leq i\leq n, 1\leq j\leq p} |x_{ij}| > K \right) \leq P\left(\max_{1\leq i\leq n, 1\leq j\leq p} |z_{ij}| > K \right) \tag{4.15}$$

where $z_{ij} \sim N(0,1)$ are i.i.d. as in Example 5. (The case with $\rho < 0$ follows by changing the signs of x_{ij} for each even j and redefining the parameter $\beta(u)$ for these new regressors; so that after the transformation we obtain the design with $\rho > 0$.) The lower bound relies only on the independence within the components of each vector $x_{\cdot j}$. Since $x_{i'j}$ and x_{ij} are independent for $i' \neq i$, we can invoke the same results of Example 5. Therefore we obtain $c_{n,p} \simeq_P \sqrt{\frac{\log(n\vee p)}{n}}$ and $t^2 \log^2(n \vee p) = o(n)$. In addition, $\phi(1) \geq 1$ and $\phi(m_0 + s) \lesssim_P \{(1+|\rho|)/(1-|\rho|)\} \left(1 + \sqrt{(m_0/n)\log p} + \sqrt{(s/n)\log p}\right) \lesssim \{(1+|\rho|)/(1-|\rho|)\}$ by Lemma 14 and the definition of m_0. Since ρ is fixed it follows that $\phi(1) \simeq_P \phi(m_0 + s)$. Thus, Theorem 4 also requires $t^2 s \log(n \vee p) = o(n)$ in this case. We also verify that $q \simeq \mu \simeq 1$ in the next section.

5. Empirical Performance. In order to access the finite sample practical performance of the proposed estimators, we conducted a Monte Carlo study and an application to international economic growth.

5.1. *Monte Carlo Simulation.* In this section we will compare the performance of the canonical quantile regression estimator, the ℓ_1-penalized quantile regression, the two-step estimator, and the ideal oracle estimator. Recall that the two-step estimator applies the canonical quantile regression to the model selected by the penalized estimator. The oracle estimator applies the canonical quantile regression on the minimal true model. (Of course, such an estimator is not available outside Monte Carlo experiments.) We focus our attention on model selection properties of the penalized estimator and biases and standard deviations of these estimators.

We begin by considering the following regression model (see Example 1) where

$$y = x'\beta(1/2) + \varepsilon, \quad \beta(1/2) = (1, 1, 1, 1, 1, 0, \ldots, 0)',$$

where an intercept and the covariates $x_{-1} \sim N(0, I)$, and the errors ε are independent identically distributed $\varepsilon \sim N(0, 1)$. We set the dimension p of covariates x equal to 1000, and the dimension s of the true minimal model to 5, and the sample size n to 200. We set the regularization parameter λ equal to 0.9-quantile of the pivotal random variable Λ_n, following our proposal in Section 2.

We also consider a variant of the model above with correlated regressors, namely $x_{-1} \sim N(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$, as specified in Example 2 with $\rho = 0.5$. This design is noteworthy because it violates the condition of the uniform uncertainty principle, but it easily satisfies our conditions.

We summarize the results on model selection performance of the penalized estimator in Figures 2-3. In the left panels of Figures 2-3, we plot the frequencies of the dimensions of the selected model; in the right panel we plot the frequencies of selecting the correct components. From the right panels we see that the model selection performance is particularly good. From the left panels we see that the frequency of selecting much larger model than the minimal true model is very small. We also see that in the design with correlated regressors, the performance of the estimator is quite good, as we would expect from our theoretical results. These results confirm the theoretical results of Theorem 2, namely, that when the non-zero coefficients are well-separated from zero, with probability tending to one, the penalized estimator should select the model that includes the minimal true model as a subset. Moreover, these results also confirm the theoretical result of Theorem 1, namely that the dimension of the selected model should be of the same stochastic order as the dimension of the true minimal model. In summary, we that find the model selection performance of the penalized estimator very well agree with our theoretical results.

We summarize results on the estimation performance in Table 1. We see that the penalized quantile regression estimator significantly outperforms the canonical quantile regression, as we would expect from Theorem 1 and from inconsistency of the latter when the number of regressors is larger than the sample size. The penalized quantile regression has a substantial bias, as we would expect from the definition of the estimator which penalizes large deviations of coefficients from zero. Furthermore, we see that the two-step estimator improves upon the penalized quantile regression, particularly in terms of drastically reducing the bias. The two-step estimator in fact does almost as well as the ideal oracle estimator, as we would expect from Theorem 4. We also see that the (unarbitrary) correlation of regressors does not harm the performance of the penalized and the two-step estimators, which we would expect from our theoretical results. In fact, since data-driven value of λ tends to be slightly lower for the correlated case, as we would expect by the comparison theorem mentioned in Example 8, the penalized estimator selects smaller models and also makes smaller estimation errors than in the canonical uncorrelated case. In summary, we find the estimation performance of the penalized and two-step estimators to be in agreement with our theoretical results.

MONTE CARLO RESULTS

Example 1: Isotropic Gaussian Design

	Mean ℓ_0 norm	Mean ℓ_1 norm	Bias	Std Deviation
Canonical QR	992.29	25.27	1.6929	0.99
Penalized QR	5.14	2.43	1.1519	0.37
Two-step QR	5.14	4.97	0.0276	0.29
Oracle QR	5.00	5.00	0.0012	0.20

Example 2: Correlated Gaussian Design

	Mean ℓ_0 norm	Mean ℓ_1 norm	Bias	Std Deviation
Canonical QR	988.41	29.40	1.2526	1.11
Penalized QR	5.19	4.09	0.4316	0.29
Two-step QR	5.19	5.02	0.0075	0.27
Oracle QR	5.00	5.00	0.0013	0.25

TABLE 1

The table displays the average ℓ_0 and ℓ_1 norm of the estimators as well as mean bias and standard deviation. We obtained the results using 5000 Monte Carlo repetitions for each design.

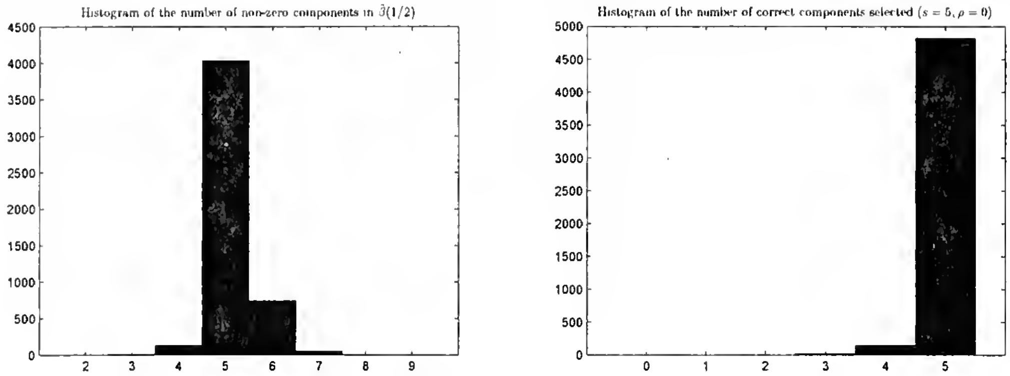

FIG 2. *The figure summarizes the covariate selection results for the isotropic normal design example, based on* 5000 *Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible 1000 covariates. The right panel plots the histogram for the number of significant covariates selected; there are in total* 5 *significant covariates amongst* 1000 *covariates.*

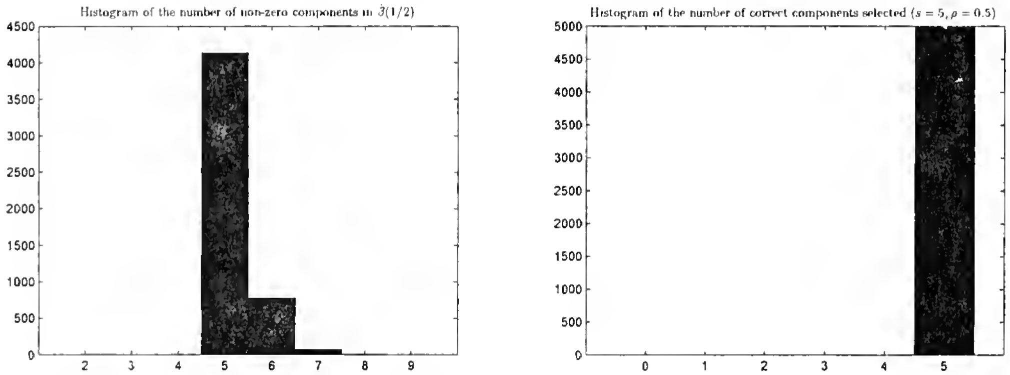

FIG 3. *The figure summarizes the covariate selection results for the correlated normal design example with correlation coefficient* $\rho = .5$, *based on* 5000 *Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible 1000 covariates. The right panel plots the histogram for the number of significant covariates selected; there are in total* 5 *significant covariates amongst* 1000 *covariates. We obtained the results using* 5000 *Monte Carlo repetitions.*

5.2. *International Economic Growth Example.* In this section we apply ℓ_1-penalized quantile regression to an international economic growth example, using it primarily as a method for model selection. We use the Barro and Lee data consisting of a panel of 138 countries for the period of 1960 to 1985. We consider the national growth rates in gross domestic product (GDP) per capita as a dependent variable y for periods 1965-75 and 1975-85.[1] In our analysis, we will consider model with nearly $p = 60$ covariates, which allows for a total of $n = 90$ complete observations. Our goal here is to select a subset of these covariates and briefly compare the resulting models to the standard models used in the empirical growth literature (Barro and Sala-i-Martin [2], Koenker and Machado [21]).

One of the central issues in the empirical growth literature is the estimation of the effect of an initial (lagged) level of GDP per capita on the growth rates of GDP per capita. In particular, a key prediction from the classical Solow-Swan-Ramsey growth model is the hypothesis of convergence, which states that poorer countries should typically grow faster and therefore should tend to catch up with the richer countries. Thus, such a hypothesis states that the effect of initial level of GDP on the growth rate should be negative. As pointed out in Barro and Sala-i-Martin [3], this hypothesis is rejected using a simple bivariate regression of growth rates on the initial level of GDP. (In our case, median regression yields a positive coefficient of 0.00045.) In order to reconcile the data and the theory, the literature has focused on estimating the effect *conditional* on the pertinent characteristics of countries. Covariates that describe such characteristics can include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others [3]. The theory then predicts that for countries with similar other characteristics the effect of the initial level of GDP on the growth rate should be negative ([3])

Given that the number of covariates we can condition on is comparable to the sample size, the covariate selection becomes an important issue in this analysis ([24], [31]). In particular, past previous findings came under severe criticisms for relying upon ad hoc procedures for covariate selection. In fact, in some cases, all of the previous findings have been questioned ([24]). Since the number of covariates is high, there is no simple way to resolve the model selection problem using only the classical tools. Indeed the number of possible lower-dimensional model is very large, though see [24] and [31] for an attempt to search over several millions of these models. Here we use the lasso selection device, specifically the ℓ_1-penalized median regressions, to resolve this important issue.

[1] The growth rate in GDP over period from t_1 to t_2 is commonly defined as $\log(GDP_{t_2}/GDP_{t_1}) - 1$.

Let us now turn to our empirical results. We performed covariate selection using the ℓ_1-penalized median regressions, where we initially used our data-driven choice of penalization parameter λ. This initial choice led us to select no covariates, which is consistent with the situations in which the true coefficients are not well-separated from zero. We then proceeded to slowly decrease the penalization parameter in order to allow for some covariates to be selected. We present the model selection results in Table 2. With the first relaxation of the choice of λ, we select the black market exchange rate premium (characterizing trade openness) and a measure of political instability. With a second relaxation of the choice of λ we select an additional set of educational attainment variables, and several others reported in the table. With a third relaxation of λ we include yet another set of variables also reported in the table. We refer the reader to [2] and [3] for a complete definition and discussion of each of these variables.

We then proceeded to apply the standard median and quantile regressions to the selected models and we also report the standard confidence intervals for these estimates. In Figures 4 and 5 we show these results graphically, plotting estimates of quantile regression coefficients $\widehat{\beta}(u)$ and pointwise confidence intervals on the vertical axis against the quantile index u on the horizontal axis. We should note that the confidence intervals do not take into account that we have selected the models using the data. (In an ongoing companion work, we are working on devising procedures that will account for this.) We find that, in all models that we have selected, the median regression coefficients on the initial level of GDP is always negative and the standard confidence intervals do not include zero. Similar conclusions also hold for quantile regressions with quantile indices in the middle range. In summary, we believe that our empirical findings support the hypothesis of convergence from the classical Solow-Swan-Ramsey growth model. Of course, it would be good to find formal inferential methods to fully support this hypothesis. Finally, our findings also agree and thus support the previous findings reported in Barro and Sala-i-Martin [2] and Koenker and Machado [21].

6. Conclusion and Extensions. In this work we characterize the estimation and model selection properties of the ℓ_1-penalized quantile regression for high-dimensional sparse models. Despite the non-linear nature of the problem, we provide results on estimation and model selection that parallel those recently obtained for the penalized least squares estimator. It is likely that our proof techniques can be useful for deriving results for other M-estimation problems.

MODEL SELECTION RESULTS FOR THE INTERNATIONAL GROWTH REGRESSIONS

Penalization Parameter $\lambda = 1.077968$	Real GDP per capita (log) is included in all models Additional Selected Variables
λ	-
$\lambda/2$	Black Market Premium (log)
	Political Instability
$\lambda/3$	Black Market Premium (log)
	Political Instability
	Measure of tariff restriction
	Infant mortality rate
	Ratio of real government “consumption” net of defense and education
	Exchange rate
	% of “higher school complete” in female population
	% of “secondary school complete” in male population
$\lambda/4$	Black Market Premium (log)
	Political Instability
	Measure of tariff restriction
	Infant mortality rate
	Ratio of real government “consumption” net of defense and education
	Exchange rate
	% of “higher school complete” in female population
	% of “secondary school complete” in male population
	Female gross enrollment ratio for higher education
	% of “no education” in the male population
	Population proportion over 65
	Average years of secondary schooling in the male population
$\lambda/5$	Black Market Premium (log)
	Political Instability
	Measure of tariff restriction
	Infant mortality rate
	Ratio of real government “consumption” net of defense and education
	Exchange rate
	% of “higher school complete” in female population
	% of “secondary school complete” in male population
	Female gross enrollment ratio for higher education
	% of “no education” in the male population
	Population proportion over 65
	Average years of secondary schooling in the male population
	Growth rate of population
	% of “higher school attained” in male population
	Ratio of nominal government expenditure on defense to nominal GDP
	Ratio of import to GDP

TABLE 2

For this particular decreasing sequence of penalization parameters we obtained nested models. All the columns of the design matrix were normalized to have unit length.

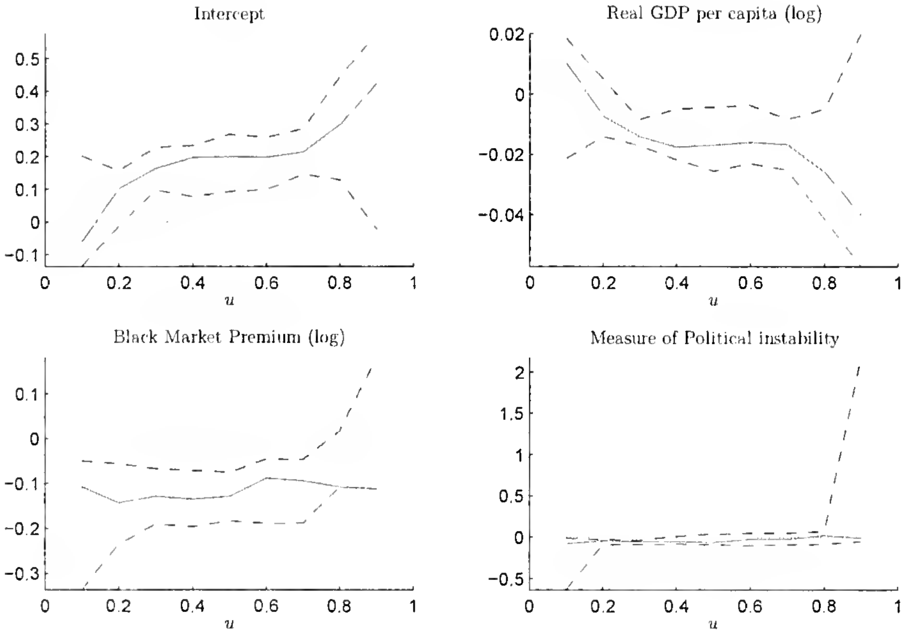

FIG 4. *This figure plots the coefficient estimates and standard pointwise 90 % confidence intervals for the model associated with $\lambda/2$ which selected two covariates in addition to the initial level of GDP.*

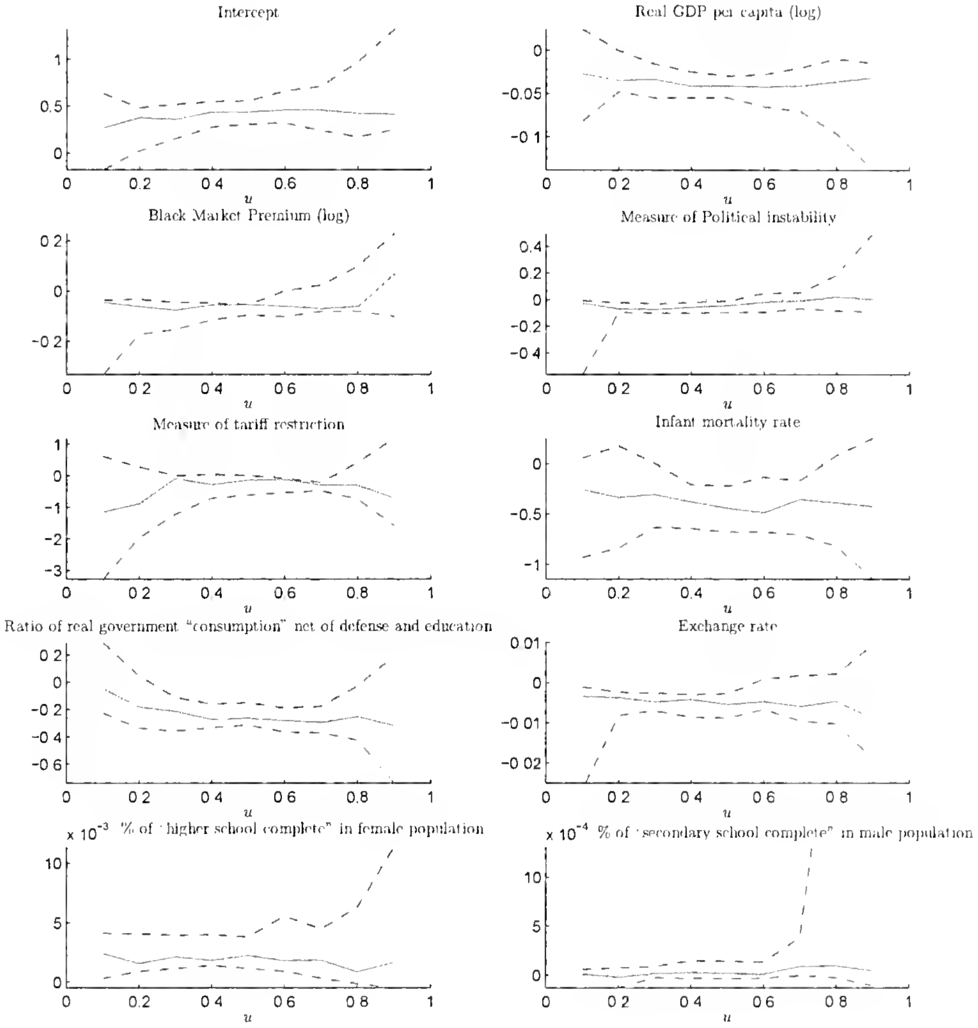

FIG 5. *This figure plots the coefficient estimates and standard pointwise 90% confidence intervals for the model associated with $\lambda/3$ which selected eight covariates in addition to the initial level of GDP.*

There are several possible extensions that we would like to pursue in the future work. First, we would like to extend out results to hold uniformly across a continuum of quantile indices. We expect that most of our results will generalize to this case with a few appropriate modifications. Second, following van der Geer [37], we would like to allow for regressor-specific choice of the penalization parameter. Specifically, we would like to consider the following estimator:

$$\widehat{\beta}(u) \in \arg\min_{\beta \in \mathbb{R}^p} \ \mathbb{E}_n\left[\rho_u(y_i - x_i'\beta)\right] + \frac{\lambda}{n}\sum_{j=1}^{p} \widehat{\sigma}_j |\beta_j| \tag{6.1}$$

where $\widehat{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij}^2}$. The dual problem associated with (6.1) has the form:

$$\begin{array}{ll} \max\limits_{a \in \mathbb{R}^n} & \mathbb{E}_n\left[y_i a_i\right] \\ & |\mathbb{E}_n\left[x_{ij} a_i\right]| \leq \frac{\lambda}{n}\widehat{\sigma}_j, \quad j = 1, \ldots, p, \\ & (u-1) \leq a_i \leq u, \quad i = 1, \ldots, n. \end{array} \tag{6.2}$$

To map this to our previous framework, we can redefine the regressors and the parameter spaces via transformations $\tilde{x}_{ij} = x_{ij}/\widehat{\sigma}_j$ and $\tilde{\beta}_j(u) = \widehat{\sigma}_j\beta_j(u)$. We can then proceed with an analogous proof strategy. Third, we would like to extend our analysis to cover non-sparse models that are well-approximated by sparse models. In such a framework, the components of $\beta(u)$ reordered by magnitude, namely $|\beta_{(1)}(u)| \geq |\beta_{(2)}(u)| \geq \cdots \geq |\beta_{(p-1)}(u)| \geq |\beta_{(p)}(u)|$, exhibit a sufficiently rapid decay behavior, for example, $|\beta_{(k)}(u)| < Rk^{-1/t}$ for some constants R and t. Therefore, truncation to zero of all components below a particular moving threshold can still lead to consistent estimation.

APPENDIX A: VERIFICATION OF CONDITIONS E.1-E.5

In this section we verify that conditions E.1-E.5 hold under the simple set of conditions D.1-D.4 discussed in Section 2 and also hold much more generally. For convenience, we denote by

$$\mathbb{S}_p^k = \{\alpha \in \mathbb{R}^p : \|\alpha\| = 1, \|\alpha\|_0 \leq k\}$$

the k-sparse unit sphere in $\mathbb{R}^p$. In what follows, we show how the key constants, such as q and μ appearing in E.1-E.5, are functions of the following population constants (which can possibly depend on the sample size n):

$$\begin{aligned} \underline{f} &:= \inf_{x} \ f_{y_i|x_i}(x'\beta(u)|x), & \bar{f} &:= \sup_{y,x} \ f_{y_i|x_i}(y|x), \\ \varrho(k) &:= \inf_{\alpha \in \mathbb{S}_p^k} \ \mathrm{E}\left[(\alpha' x)^2\right], & \bar{f}' &:= \sup_{y,x} \ \frac{\partial}{\partial y} f_{y_i|x_i}(y|x), \\ \gamma(k) &:= \inf_{\alpha \in \mathbb{S}_p^k} \ \frac{\mathrm{E}\left[|\alpha' x|^2\right]}{\mathrm{E}\left[|\alpha' x|^3\right]}, & \varphi(k) &:= \sup_{\alpha \in \mathbb{S}_p^k} \ \mathrm{E}\left[(\alpha' x_i)^2\right], \end{aligned} \tag{A.1}$$

where values of y and x range over the support of y_i and x_i. The results also depend on the sparse eigenvalue of sample design matrix

$$\phi(k) := \sup_{\alpha \in \mathbb{S}_p^k} \mathbb{E}_n\left[(\alpha' x_i)^2\right]$$

already mentioned earlier. As an illustration we compute the constants in (A.1) for two common designs used in the literature.

Example 7 (Isotropic Normal Design, continued). We revisit the design of Example 1. For concreteness assume that the errors are $\varepsilon \sim N(0,1)$. Under this simple design we can compute the values of the several constants involved in the analysis: $\bar{f} = 1/\sqrt{2\pi} \leq 0.4$, $\underline{f} = 1/\sqrt{2\pi} \geq 0.39$, $\bar{f}' = 1/\sqrt{2\pi e} \leq 0.25$, $\gamma(k) = \sqrt{\pi/8} \geq 0.6$, $\varrho(k) = 1$, and $\varphi(k) = 1$.

Example 8 (Correlated Normal Design, continued). Consider next the design in Example 2. For concreteness assume that $\varepsilon \sim N(0,1)$ and that $\rho = 1/2$. The relevant constants are bounded by $\bar{f} = 1/\sqrt{2\pi} \leq 0.4$, $\underline{f} = 1/\sqrt{2\pi} \geq 0.39$, $\bar{f}' = 1/\sqrt{2\pi e} \leq 0.25$, $\gamma(k) \geq \sqrt{\frac{1-|\rho|}{1+|\rho|}}\sqrt{\pi/8} \geq 1/3$, $\varrho(k) \geq \frac{1}{2}\frac{1-|\rho|}{1+|\rho|} = 1/6$, and $\varphi(k) \leq \frac{1+|\rho|}{1-|\rho|} = 3$.

A.1. Verification of E.1-E.5. Conditions E.1 (model sparseness) is the key underlying model assumption, which we impose throughout, including in condition D.2. Lemmas 1, 2, 3, 4, and 5 below establish the remaining conditions E.2-E.5.

Lemma 1 (Verifying Condition E.2 - Identification). *We have that in the linear quantile model (2.1) under random sampling, for each $\beta \in \mathbb{R}^p : \|\beta - \beta(u)\| = r, \|\beta - \beta(u)\|_0 \leq m$,*

$$Q_u(\beta) - Q_u(\beta(u)) \geq q(m)\left(r^2 \wedge g(r)\right), \tag{A.2}$$

where

$$g(r) = r \quad \textit{and} \quad q(m) = \frac{\varrho(m)\underline{f}}{4}\min\left\{1, \left(\frac{3\underline{f}}{2\bar{f}'}\gamma(m)\right)^2\right\}. \tag{A.3}$$

Thus, condition E.2 holds with $q = q(n)$. In particular, under Conditions D.1-D.4, condition E.2 holds with $q = q(n) \simeq 1$.

PROOF. (Lemma 1) Let $F_{y|x}$ denote the conditional distribution of y given x. From Knight [17], for any two scalars w and v we have that

$$\rho_u(w-v) - \rho_u(w) = -v(u - 1\{w \leq 0\}) + \int_0^v (1\{w \leq z\} - 1\{w \leq 0\})dz. \tag{A.4}$$

Applying (A.4) with $w = y - x'\beta(u)$ and $v = x'(\beta - \beta(u))$ we have that $\mathrm{E}\left[-v(u - 1\{w \leq 0\})\right] = 0$. Using the law of iterated expectations and mean value expansion, we obtain for $\tilde{z}_{x,z} \in [0, z]$

$$\begin{array}{rcl} Q_u(\beta) - Q_u(\beta(u)) & = & \mathrm{E}\left[\int_0^{x'(\beta-\beta(u))} F_{y|x}(x'\beta(u) + z) - F_{y|x}(x'\beta(u))dz\right] \\ & = & \mathrm{E}\left[\int_0^{x'(\beta-\beta(u))} z f_{y|x}(x'\beta(u)) + \frac{z^2}{2} f'_{y|x}(x'\beta(u) + \tilde{z}_{x,z})dz\right] \\ & \geq & \mathrm{E}\left[\frac{1}{2}(x'(\beta-\beta(u)))^2 f_{y|x}(x'\beta(u))\right] - \frac{\bar{f}'}{6}\mathrm{E}[|x_i'(\beta-\beta(u))|^3] \\ & \geq & \frac{\underline{f}}{2}\mathrm{E}\left[(x'(\beta-\beta(u)))^2\right] - \frac{\bar{f}'}{6}\mathrm{E}[|x'(\beta-\beta(u))|^3]. \end{array} \tag{A.5}$$

Next define

$$r_m = \sup\left\{\tilde{r} \ : \ Q_u(\beta(u) + \tilde{r}d) - Q_u(\beta(u)) \geq \frac{\underline{f}}{4}\tilde{r}^2 \mathrm{E}\left[(x'd)^2\right], \text{ for all } d \in \mathbb{S}_p^m\right\}.$$

By (A.5) we have that

$$r_m \geq \frac{3}{2}\frac{\underline{f}}{\bar{f}'} \inf_{\alpha \in \mathbb{S}_p^m} \frac{\mathrm{E}\left[|\alpha' x|^2\right]}{\mathrm{E}\left[|\alpha' x|^3\right]} = \frac{3}{2}\frac{\underline{f}}{\bar{f}'}\gamma(m).$$

By construction of r_m and the convexity of Q_u, for any β such that $\|\beta - \beta(u)\| \leq m$, we have that

$$Q_u(\beta) - Q_u(\beta(u)) \geq \frac{\underline{f}}{4}\mathrm{E}\left[(x'(\beta-\beta(u)))^2\right] \wedge \left\{\|\beta - \beta(u)\| \left(\inf_{d \in \mathbb{S}_p^m} Q_u(\beta(u) + r_m d) - Q_u(\beta(u))\right)\right\}.$$

Letting $\|\beta - \beta(u)\| = r$ we have

$$r\left(\inf_{d \in \mathbb{S}_p^m} Q_u(\beta(u) + r_m d) - Q_u(\beta(u))\right) \geq \frac{r\underline{f}\varrho(m) r_m^2}{4} \text{ and } \frac{\underline{f}}{4}\mathrm{E}\left[(x'(\beta-\beta(u)))^2\right] \geq \frac{r^2 \underline{f}\varrho(m)}{4},$$

where the first inequality holds by construction of r_m; hence

$$Q_u(\beta) - Q_u(\beta(u)) \geq \frac{r^2\underline{f}\varrho(m)}{4} \wedge \frac{r\underline{f}\varrho(m)r_m^2}{4} \geq q(m)(r^2 \wedge r)$$

for $q(m)$ defined in (A.3). □

The following two lemmas verify the Empirical Pre-sparseness condition.

LEMMA 2 (Verifying Condition E.3 - Empirical Pre-Sparseness). *We have that the number of non-zero components of $\widehat{\beta}(u)$ is bounded by $n \wedge p$, namely*

$$\|\widehat{\beta}(u)\|_0 \leq n \wedge p.$$

Suppose that $y_1, \ldots, y_n$ are absolutely continuous conditional on $x_1, \ldots, x_n$, then the number of interpolated points, $h = |\{i : y_i = x_i'\widehat{\beta}(u)\}|$, is equal to $\|\widehat{\beta}(u)\|_0$ with probability one.

PROOF. Trivially we have $\|\widehat{\beta}(u)\|_0 \leq p$. Let $Y = (y_1, \ldots, y_n)'$, X be the $n \times p$ matrix with rows $x_i', i = 1, \ldots, n$, $c = (ue', (1-u)e', \lambda e', \lambda e')'$, and $A = [I \ \ -I \ \ X \ \ -X]$, where $e = (1, 1, \ldots, 1)'$ denotes vectors of ones of conformable dimensions, and I denotes the $n \times n$ identity matrix. Note that the penalized quantile regression can be written as

$$\begin{array}{cl} \displaystyle\min_{\xi^+,\xi^-,\beta^+,\beta^-} & ue'\xi^+ + (1-u)e'\xi^- + \lambda e'\beta^+ + \lambda e'\beta^- \\ & \xi^+ - \xi^- + X\beta^+ - X\beta^- = Y \\ & (\xi^+, \xi^-, \beta^+, \beta^-) \in \mathbb{R}_+^{2n+2p} \end{array} \quad \Leftrightarrow \quad \begin{array}{cl} \displaystyle\min_{w} & c'w \\ & Aw = Y \\ & w \geq 0. \end{array}$$

Matrix A has rank n, since it has linearly independent rows. By Theorem 2.4 of Bertsimas and Tsitsiklis [6] there is at least one optimal basic solution w^* with at most n non-zero components. We defined $\widehat{\beta}(u)$ as a basic solution with the minimal number of non-zero components (note that $\|\widehat{\beta}(u)\|_0 = \|\widehat{\beta}^+(u)\|_0 + \|\widehat{\beta}^-(u)\|_0$ since $\lambda > 0$). Let h denote the number of interpolated points. We have that $n-h$ components of ξ and $\bar{\xi}$ are non-zero. Therefore, we have $\|\widehat{\beta}(u)\|_0 + (n-h) \leq n$ which leads to $\|\widehat{\beta}(u)\|_0 \leq h \leq n$.

To prove the second statement, consider the dual problem $\max_a\{Y'a : A'a \leq c\}$. Conditional on X consider the polyhedron defined by $\{a : A'a \leq c\}$ which has a finite number of vertices. Since $c > 0$ componentwise this polyhedron is non-empty (i.e., zero is always feasible for the dual problem). Moreover, the form of A' implies that $\{a : A'a \leq c\}$ is a bounded set. Therefore, if the solution of the dual is not unique there exist vertices a^1, a^2 such that $Y'(a^1 - a^2) = 0$. This is a zero probability event since Y is absolutely continuous conditional on X and the number of vertices is finite. Therefore the dual problem has a unique solution with probability one. If the dual basic solution is unique, we have that the primal basic solution is non-degenerate, that is, the number of non-zero variables equals n, see [6]. Therefore, we have with probability one that $\|\widehat{\beta}(u)\|_0 + (n-h) = n$, or that $\|\widehat{\beta}(u)\|_0 = h$. □

From the complementary slackness condition of linear programming, see Theorem 4.5 of [6], we have that for any component $j \in \{1, \ldots, p\}$

$$\text{(A.6)} \qquad \begin{array}{lll} \widehat{\beta}_j(u) > 0 & \text{only if} & \mathbb{E}_n\left[x_{ij}\widehat{a}_i(u)\right] = \frac{\lambda}{n}, \text{ and} \\ \widehat{\beta}_j(u) < 0 & \text{only if} & \mathbb{E}_n\left[x_{ij}\widehat{a}_i(u)\right] = -\frac{\lambda}{n} \end{array}$$

where $\widehat{a}(u)$ solves the dual problem (2.6).

LEMMA 3 (Verifying Condition E.3 - Empirical Pre-Sparseness, continued). *Let $m = \|\widehat{\beta}(u)\|_0$. For any $\lambda > 0$ we have*

$$m \leq \frac{n^2\phi(m)}{\lambda^2}.$$

PROOF. Let $\widehat{a}(u)$ be the solution of the dual problem (2.6), $\widehat{T} =$ support$(\widehat{\beta}(u))$, and $m = \|\widehat{\beta}(u)\|_0 = |\widehat{T}|$. For any $k \in \widehat{T}$, from (A.6) we have $(X'\widehat{a}(u))_k = \text{sign}(\widehat{\beta}_k(u))\lambda$ and, for $k \notin \widehat{T}$ we have $\text{sign}(\widehat{\beta}_k(u)) = 0$. Therefore, by Cauchy-Schwarz inequality we have

$$\begin{aligned} m\lambda &= \text{sign}(\widehat{\beta}(u))'\text{sign}(\widehat{\beta}(u))\lambda = \text{sign}(\widehat{\beta}(u))'(X'\widehat{a}(u)) = \left(X\text{sign}(\widehat{\beta}(u))\right)'\widehat{a}(u) \\ &\leq \|X\text{sign}(\widehat{\beta}(u))\|\|\widehat{a}(u)\| \leq \sqrt{n\phi(m)}\|\text{sign}(\widehat{\beta}(u))\|\|\widehat{a}(u)\|, \end{aligned}$$

where we used that $\|\text{sign}(\widehat{\beta}(u))\|_0 = m$. Since $\|\widehat{a}(u)\| \leq \sqrt{\max\{u, 1-u\}n} \leq \sqrt{n}$, and $\|\text{sign}(\widehat{\beta}(u))\| = \sqrt{m}$ we have $m\lambda \leq n\sqrt{m\phi(m)}$, which yields the result. □

We shall need some additional notation in what follows. Let

$$\psi_i(\beta, u) = (1\{y_i \leq x_i'\beta\} - u)x_i$$

denote the score function for the ith observation. Define the set of m-sparse vectors near to the true value $\beta(u)$

$$R(r, m) := \{\beta \in \mathbb{R}^p \ : \|\beta\|_0 \leq m, \ \|\beta - \beta(u)\| \leq r\},$$

and define the sparse sphere associated with a given vector β as

$$\mathbb{S}(\beta) = \{\alpha \in \mathbb{R}^p : \|\alpha\| \leq 1, \text{support}(\alpha) \subseteq \text{support}(\beta)\}.$$

Also, define the following empirical and linearization errors

$$\text{(A.7)} \qquad \begin{array}{lll} \epsilon_0(m, n, p) & := \sup_{\alpha \in \mathbb{S}_p^m} & |\mathbb{G}_n(\alpha'\psi_i(\beta(u), u))|, \\ \epsilon_1(r, m, n, p) & := \sup_{\beta \in R(r,m), \alpha \in \mathbb{S}(\beta)} & |\mathbb{G}_n(\alpha'\psi_i(\beta, u)) - \mathbb{G}_n(\alpha'\psi_i(\beta(u), u))|, \\ \epsilon_2(r, m, n, p) & := \sup_{\beta \in R(r,m), \alpha \in \mathbb{S}(\beta)} & \sqrt{n}|\mathrm{E}[\alpha'\psi_i(\beta, u)] - \mathrm{E}[\alpha'\psi_i(\beta(u), u)]|. \end{array}$$

where $\mathbb{G}_n$ is the empirical process operator, that is $\mathbb{G}_n(f) := n^{-1/2}\sum_{i=1}^n (f(X_i) - \mathrm{E}[f(X_i)])$.

Next we verify condition E.4.

Lemma 4 (Verifying Condition E.4 - Empirical Sparseness). *Let $m = \|\widehat{\beta}(u)\|_0$, $r = \|\widehat{\beta}(u) - \beta(u)\|$, and suppose that $y_1, \ldots, y_n$ are absolutely continuous conditional on $x_1, \ldots, x_n$. We have that, in the linear quantile model (2.1) under random sampling,*

$$\sqrt{m} \lesssim_p \mu(m)\ \frac{n}{\lambda}(r \wedge 1) + \sqrt{m}\frac{\left(\sqrt{n\log(n \vee p)\phi(m)} \vee \sqrt{n\log(n\vee p)\varphi(m)}\right)}{\lambda},$$

uniformly in $m \leq n$ and r, where $\mu(m) = \sqrt{\varphi(m)}(\sqrt{\varphi(m)}\bar{f} \vee 1)$. Therefore, provided $\varphi(m) \lesssim_p \phi(m)$, condition E.4 holds with $\mu = \mu(n)$, namely

$$\sqrt{m} \lesssim_p \mu\ \frac{n}{\lambda}(r \wedge 1) + \sqrt{m}\frac{\sqrt{n\log(n\vee p)\phi(m)}}{\lambda}.$$

In particular, under D.1-D.4, $\varphi(m) \lesssim 1$ and $\phi(1) \geq 1$, so that condition E.4 holds with $\mu = \mu(n) \simeq 1$.

Proof. (Lemma 4) It will be convenient to define three vectors of rank scores (dual variables):

1. the true rank scores, $a_i^*(u) = u - 1\{y_i \leq x_i'\beta(u)\}$ for $i = 1, \ldots, n$;
2. the estimated rank scores, $a_i(u) = u - 1\{y_i \leq x_i'\widehat{\beta}(u)\}$ for $i = 1, \ldots, n$;
3. the dual optimal rank scores, $\widehat{a}(u)$ that solve the dual program (2.6).

Let $\widehat{T}$ denote the support of $\widehat{\beta}(u)$. Let $x_{i\widehat{T}} = (x_{ij}, j \in \widehat{T})'$, and $\widehat{\beta}_{\widehat{T}}(u) = (\widehat{\beta}_j(u), j \in \widehat{T})'$. From the complementary slackness characterizations (A.6) we have that

$$\text{(A.8)}\qquad \mathrm{sign}(\widehat{\beta}_{\widehat{T}}(u)) = \frac{n\mathbb{E}_n\left[x_{i\widehat{T}}\widehat{a}_i(u)\right]}{\lambda}, \text{ i.e. } \sqrt{m} = \|\mathrm{sign}(\widehat{\beta}_{\widehat{T}}(u))\| = \left\|\frac{n\mathbb{E}_n\left[x_{i\widehat{T}}\widehat{a}_i(u)\right]}{\lambda}\right\|.$$

Therefore we can bound the number of non-zero components of $\widehat{\beta}(u)$ provided we can bound the empirical expectation in (A.8). This is achieved in the next step by combining the maximal inequalities and assumptions on the design matrix.

Using the triangle inequality in (A.8), write

$$\lambda\sqrt{m} \leq \left\|n\mathbb{E}_n\left[x_{i\widehat{T}}(\widehat{a}_i(u) - a_i(u))\right]\right\| + \left\|n\mathbb{E}_n\left[x_{i\widehat{T}}(a_i(u) - a_i^*(u))\right]\right\| + \left\|n\mathbb{E}_n\left[x_{i\widehat{T}}a_i^*(u)\right]\right\|.$$

Then we bound each of the three components in this display. To bound the last component, we use Lemma 9 to get

$$\left\|n\mathbb{E}_n\left[x_{i\widehat{T}}a_i^*(u)\right]\right\| \leq \sqrt{n}\epsilon_0(m,n,p) \lesssim_P \sqrt{nm\log(n \vee p)}\left(\sqrt{\varphi(m)} \vee \sqrt{\phi(m)}\right).$$

To bound the first component, we observe that $\widehat{a}_i(u) \neq a_i(u)$ only if $y_i = x_i'\widehat{\beta}(u)$. By Lemma 2 the penalized quantile regression fit can interpolate at most m points with probability one. This implies that $\mathbb{E}_n\left[|\widehat{a}_i(u) - a_i(u)|^2\right] \leq m/n$. Therefore, we get

$$\begin{aligned}\left\|n\mathbb{E}_n\left[x_{i\widehat{T}}(\widehat{a}_i(u) - a_i(u))\right]\right\| &\leq n\sup_{\alpha \in \mathbb{S}_p^m}\mathbb{E}_n\left[|\alpha' x_i|\ |\widehat{a}_i(u) - a_i(u)|\right] \\ &\leq n\sup_{\alpha\in\mathbb{S}_p^m}\sqrt{\mathbb{E}_n\left[|\alpha' x_i|^2\right]}\sqrt{\mathbb{E}_n\left[|\widehat{a}_i(u) - a_i(u)|^2\right]} \\ &\leq \sqrt{n\phi(m)m}.\end{aligned}$$

To bound the second component, note that

$$\begin{aligned}\left\|n\mathbb{E}_n\left[x_{i\widehat{T}}(a_i(u) - a_i^*(u))\right]\right\| &= \left\|\sqrt{n}\ \mathbb{G}_n\left(x_{i\widehat{T}}(a_i(u) - a_i^*(u))\right)\right\| + \left\|n\mathrm{E}\left[x_{i\widehat{T}}(a_i(u) - a_i^*(u))\right]\right\| \\ &\leq \sqrt{n}\epsilon_1(r,m,n,p) + \sqrt{n}\epsilon_2(r,m,n,p) \\ &\lesssim_P \sqrt{nm\log(n\vee p)}\sqrt{\varphi(m)\vee\phi(m)} + n\sqrt{\varphi(m)}(\sqrt{\varphi(m)}\bar{f}r \wedge 1)\end{aligned}$$

where we use Lemma 8 and Lemma 10 to bound respectively $\epsilon_1(r,m,n,p)$ and $\epsilon_2(r,m,n,p)$.

Setting $\mu(m) = \sqrt{\varphi(m)}(\sqrt{\varphi(m)}\bar{f} \vee 1) \geq \sqrt{\varphi(m)}$ and using that $m \leq n$ the first result follows.

Under D.1-D.4 we $\varphi(n) \lesssim 1$, and $\phi(1) \geq 1$ and condition E.4 holds. □

LEMMA 5 (Verifying Condition E.5 - Empirical Error). *We have that, in the linear quantile model (2.1) under random sampling, and uniformly over $m \leq n$, $r \geq 0$, and the region $R(r,m)$:*

$$\left|\widehat{Q}_u(\beta) - Q_u(\beta) - \left(\widehat{Q}_u(\beta(u)) - Q_u(\beta(u))\right)\right| \lesssim_P \frac{r\sqrt{(m+s)\log(n\vee p)}}{\sqrt{n}}\left(\sqrt{\varphi(m+s)\vee\phi(m+s)}\right).$$

In particular, under D.1-D.4 we have that condition E.5 holds, namely

$$\left|\widehat{Q}_u(\beta) - Q_u(\beta) - \left(\widehat{Q}_u(\beta(u)) - Q_u(\beta(u))\right)\right| \lesssim_P \frac{r}{\sqrt{n}}\sqrt{(m+s)\log(n\vee p)\ \phi(m+s)}$$

uniformly over $m \leq n$, $r \geq 0$, and the region $R(r,m)$.

PROOF. For convenience let $\varepsilon_n := \left|\widehat{Q}_u(\beta) - Q_u(\beta) - \left(\widehat{Q}_u(\beta(u)) - Q_u(\beta(u))\right)\right|$. Since $r \geq \|\beta - \beta(u)\|$, and $\|\beta - \beta(u)\|_0 \leq m + s$ we have that

$$\begin{aligned}\varepsilon_n &\leq \tfrac{1}{\sqrt{n}}\left|\int_0^r \tfrac{(\beta-\beta(u))'}{r}(\mathbb{G}_n(\psi_i(\tfrac{r-z}{r}\beta + \tfrac{z}{r}\beta(u),u)) - \mathbb{G}_n(\psi_i(\beta(u),u))) dz\right| \\ &\leq \tfrac{1}{\sqrt{n}}\int_0^r \epsilon_1(r,m+s,n,p)dz = \tfrac{r}{\sqrt{n}}\epsilon_1(r,m+s,n,p).\end{aligned}$$

The first result follows from Lemma 8.

Under D.1-D.4 we have $\varphi(n) \lesssim 1$ and $\phi(1) \geq 1$ and condition E.5 holds. □

A.2. Controlling Empirical and Linearization Errors. Here we exploit the technical results of Appendix A.4 to control the empirical errors ϵ_0 and ϵ_1. These technical results provide the maximal inequalities for a collection of empirical processes indexed by submodels' dimensions $m \leq n$, which may be of some independent interest. These technical results and their usage rely on the concepts of the VC dimension and the uniform covering number for a class of functions (see, e.g., [38]).

We begin with a bound on the VC dimension of relevant functions classes.

LEMMA 6. *Consider a fixed subset $T \subset \{1, 2, \dots, p\}$, $|T| = m$. The classes of functions*

$$\mathcal{F}_T = \{\alpha'(\psi_i(\beta, u) - \psi_i(\beta(u), u)) \ : \alpha \in \mathbb{S}(\beta), \text{support}(\beta) \subseteq T\}, \ \textit{and}$$

$$\mathcal{G}_T = \{\alpha'\psi_i(\beta(u), u) \ : \ \text{support}(\alpha) \subseteq T\}$$

have their VC index bounded by cm for some universal constant c.

PROOF. We prove the result for $\mathcal{F}_T$, and we omit the proof for $\mathcal{G}_T$ as it is similar.

Consider the classes of functions $\mathcal{W} := \{x'\alpha : \text{support}(\alpha) \subseteq T\}$ and $\mathcal{V} := \{1\{y \leq x'\beta\} : \text{support}(\beta) \subseteq T\}$ (for convenience let $Z = (y, x)$). Since T is fixed and has cardinality m, their VC index is bounded by $m+2$; see, for example, van der Vaart and Wellner [38] Lemma 2.6.15. Next consider $f \in \mathcal{F}_T$ which can be written in the form $f(Z) := g(Z)(1\{h(Z) \leq 0\} - 1\{p(Z) \leq 0\})$ where $g \in \mathcal{W}$, $1\{h \leq 0\}$ and $1\{p \leq 0\} \in \mathcal{V}$. The VC index of $\mathcal{F}_T$ is by definition equal to the VC index of the class of sets $\{(Z,t) : f(Z) \leq t\}, f \in \mathcal{F}_T, t \in \mathbb{R}$. We have that

$$\begin{array}{rcl} \{(Z,t) : f(Z) \leq t\} & = & \{(Z,t) : g(Z)(1\{h(Z) \leq 0\} - 1\{p(Z) \leq 0\}) \leq t\} \\ & = & \{(Z,t) : h(Z) > 0, p(Z) > 0, \ t \geq 0\} \cup \\ & \cup & \{(Z,t) : h(Z) \leq 0, p(Z) \leq 0, \ t \geq 0\} \cup \\ & \cup & \{(Z,t) : h(Z) \leq 0, p(Z) > 0, g(Z) \leq t\} \cup \\ & \cup & \{(Z,t) : h(Z) > 0, p(Z) \leq 0, g(Z) \geq t\}. \end{array}$$

Thus each set $\{(Z,t) : f(Z) \leq t\}$ is created by taking finite unions, intersections, and complements of the basic sets $\{Z : h(Z) > 0\}$, $\{Z : p(Z) \leq 0\}$, $\{t \geq 0\}$, $\{(Z,t) : g(Z) \geq t\}$, and $\{(Z,t) : g(Z) \leq t\}$. These basic sets form VC classes, each having VC index of order m.

Therefore, the VC index of a class of sets $\{(Z,t) : f(Z) \leq t\}, f \in \mathcal{F}_T, t \in \mathbb{R}$ is of the same order as the sum of the VC indices of the set classes formed by the basic VC sets; see van der Vaart and Wellner [38] Lemma 2.6.17. □

Next we control the uniform L_2 covering numbers for function classes generated by taking the union of all m-dimensional subsets of a p-dimensional set.

LEMMA 7. *For any $m \leq n$, consider the classes of functions*

$$\mathcal{F}_{m,n,p} = \{\alpha'(\psi_i(\beta,u) - \psi_i(\beta(u),u)) : \beta \in \mathbb{R}^p, \|\beta\|_0 \leq m, \alpha \in \mathbb{S}(\beta)\} \quad \textit{and}$$

$$\mathcal{G}_{m,n,p} = \left\{\alpha'\psi_i(\beta(u),u) \ : \alpha \in \mathbb{S}_p^m\right\},$$

with envelope functions $F_{m,n,p}$ and $G_{m,n,p}$. For each $\epsilon > 0$

$$\sup_Q N(\epsilon\|F_{m,n,p}\|_{Q,2}, \mathcal{F}_{m,n,p}, L_2(Q)) \leq C\left(\frac{16e}{\epsilon}\right)^{2(cm-1)}\left(\frac{ep}{m}\right)^m$$

$$\sup_Q N(\epsilon\|G_{m,n,p}\|_{Q,2}, \mathcal{G}_{m,n,p}, L_2(Q)) \leq C\left(\frac{16e}{\epsilon}\right)^{2(cm-1)}\left(\frac{ep}{m}\right)^m$$

for some universal constants C and c.

PROOF. Let $\mathcal{F}_T$ denote a restriction of $\mathcal{F}_{m,n,p}$ for a particular choice of m non-zero components. It follows that its VC dimension is at most cm by Lemma 6. In turn this implies that the covering number of $\mathcal{F}_T$ is bounded by

$$N(\epsilon\|F_T\|_{Q,2}, \mathcal{F}_T, L_2(Q)) \leq C(cm)(16e)^{cm}\left(\frac{1}{\epsilon}\right)^{2(cm-1)},$$

where C is an universal constant, see van der Vaart and Wellner [38] Theorem 2.6.7. Since we have at most $\binom{p}{m} \leq (ep/m)^m$ different restrictions T, the total covering number is bounded according the statement of the lemma. The proof for $\mathcal{G}_{m,n,p}$ follows similarly. □

Next we proceed to control the empirical errors ϵ_0 and ϵ_1 as defined in (A.7).

LEMMA 8 (Controlling error ϵ_1). *We have that in the linear quantile model (2.1)*

$$\epsilon_1(r,m,n,p) \lesssim_P \sqrt{m\log(n \vee p)}\max\left\{\sqrt{\varphi(m)},\ \sqrt{\phi(m)}\right\}.$$

uniformly in r and $m \leq n$.

PROOF. By definition of ϵ_1 we have $\epsilon_1(r, m, n, p) \leq \sup_{f \in \mathcal{F}_{m,n,p}} |\mathbb{G}_n(f)|$. From Lemma 7 the uniform covering number of $\mathcal{F}_{n,m,p}$ is bounded by $C\,(16e/\epsilon)^{2(cm-1)}\,(ep/m)^m$. Using Lemma 18 we have that uniformly in $m \leq n$

$$\sup_{f \in \mathcal{F}_{m,n,p}} |\mathbb{G}_n(f)| \lesssim_P \sqrt{m \log(n \vee p)} \max\left\{ \sup_{f \in \mathcal{F}_{m,n,p}} \mathrm{E}[f^2]^{1/2}, \sup_{f \in \mathcal{F}_{m,n,p}} \mathbb{E}_n[f^2]^{1/2} \right\} \tag{A.9}$$

Since $|\alpha' (\psi_i(\beta, u) - \psi_i(\beta(u), u))| = |\alpha' x_i|\, |1\{y_i \leq x_i'\beta\} - 1\{y_i \leq x_i'\beta(u)\}| \leq |\alpha' x_i|$,

$$\mathbb{E}_n[f^2] \leq \mathbb{E}_n\left[|\alpha' x_i|^2\right] \leq \phi(m) \text{ and } \mathrm{E}[f^2] \leq \mathrm{E}\left[|\alpha' x_i|^2\right] \leq \varphi(m), \tag{A.10}$$

using the definition of $\phi(m)$ and $\varphi(m)$. Combining (A.10) with (A.9) we obtain the result. □

Next we bound ϵ_0 using the same tools we used to bound ϵ_1.

LEMMA 9 (Controlling empirical error ϵ_0). *In the linear quantile model (2.1) we have*

$$\epsilon_0(m, n, p) \quad \lesssim_P \quad \sqrt{m \log(n \vee p)} \max\left\{ \sqrt{\varphi(m)},\ \sqrt{\phi(m)} \right\}$$

uniformly in $m \leq n$.

PROOF. The proof is similar to the proof of Lemma 8 with $\mathcal{G}_{m,n,p}$ instead of $\mathcal{F}_{m,n,p}$. Note that for $g \in \mathcal{G}_{m,n,p}$ we have $\mathbb{E}_n[g^2] = \mathbb{E}_n[(\alpha'\psi_i(\beta(u), u))^2] = \mathbb{E}_n\left[(\alpha' x_i)^2 (1\{y_i \leq x_i'\beta(u)\} - u)^2\right]$ $\leq \mathbb{E}_n\left[(\alpha' x_i)^2\right] \leq \phi(m)$ for all $\alpha \in \mathbb{S}_p^m$. □

Alternatively we could bound ϵ_0 using Theorem 5.2.2 of Stout [34] to achieve a dependence on $\phi(1)$ instead of $\phi(m)$ by making additional assumptions on the covariates x_{ij}. Now we proceed to bound ϵ_2.

LEMMA 10 (Controlling linearization error ϵ_2). *We have that in the linear quantile model*

$$\epsilon_2(r, m, n, p) \lesssim \sqrt{n}\sqrt{\varphi(m)}\left(\sqrt{\varphi(m)}\bar{f} r \wedge 1\right)$$

uniformly in $r > 0$ and $m \leq n$.

PROOF. By definition we have

$$\begin{aligned} \epsilon_2(r, m, n, p) &= \sup\nolimits_{\beta \in R(r,m), \alpha \in \mathbb{S}(\beta)} \sqrt{n} |\mathrm{E}[\alpha'\psi_i(\beta, u)] - \mathrm{E}[\alpha'\psi_i(\beta(u), u)]| \\ &= \sup\nolimits_{\beta \in R(r,m), \alpha \in \mathbb{S}(\beta)} \sqrt{n} |\mathrm{E}[(\alpha' x_i)\,(1\{y_i \leq x_i'\beta\} - 1\{y_i < x_i'\beta(u)\})]|. \end{aligned}$$

By the Cauchy-Schwarz inequality the expression above is bounded by

$$\sqrt{n} \sup_{\alpha \in \mathbb{S}_p^m} \sqrt{\mathrm{E}[(\alpha' x_i)^2]} \sup_{\beta \in R(r,m)} \sqrt{\mathrm{E}[(1\{y_i \leq x_i'\beta\} - 1\{y_i < x_i'\beta(u)\})^2]}.$$

By definition $\varphi(m) = \sup_{\alpha \in \mathbb{S}_p^m} \mathrm{E}[(\alpha' x_i)^2]$. Next, since $\mid 1\{y_i \leq x_i'\beta\} - 1\{y_i < x_i'\beta(u)\} \mid \leq 1\{|y_i - x_i'\beta(u)| \leq |x_i'(\beta - \beta(u))|\}$, we have

$$\begin{aligned}
\mathrm{E}[(1\{y_i \leq x_i'\beta\} - 1\{y_i < x_i'\beta(u)\})^2] &= \mathrm{E}\left[|1\{y_i \leq x_i'\beta\} - 1\{y_i < x_i'\beta(u)\}|\right] \\
&\leq \mathrm{E}\left[1\{|y_i - x_i'\beta(u)| \leq |x_i'(\beta - \beta(u))|\}\right] \\
&\leq \mathrm{E}\left[\int_{-|x_i'(\beta-\beta(u))|}^{|x_i'(\beta-\beta(u))|} f_{y|x}(t + x_i'\beta(u)|x_i)dt \wedge 1\right] \\
&\leq (2\bar{f}\|\beta - \beta(u)\| \sup_{\alpha \in \mathbb{S}_p^m} \mathrm{E}\left[|\alpha' x_i|\right]) \wedge 1 \\
&\leq \left(2r\bar{f}\sqrt{\varphi(m)}\right) \wedge 1.
\end{aligned}$$

□

A.3. Lemmas on Sparse Eigenvalues. In this section we collect lemmas on the maximum k-sparse eigenvalues that are used in some of the derivations presented earlier. Recall the notation for the unit sphere $\mathbb{S}^{n-1} = \{\alpha \in \mathbb{R}^n : \|\alpha\| = 1\}$ and the k-sparse unit sphere $\mathbb{S}_p^k = \{\alpha \in \mathbb{R}^p : \|\alpha\| = 1, \|\alpha\|_0 \leq k\}$. For a matrix M, let $\phi_M(k)$ denote the maximum k-sparse eigenvalue of M, namely $\phi_M(k) = \sup\{\ \alpha' M \alpha\ : \alpha \in \mathbb{S}_p^k\ \}$.

We begin with a lemma that establishes a type of subadditivity of the maximum sparse eigenvalues as a function of the cardinality.

Lemma 11. *Let M be a semi-definite positive matrix. For any integers k and ℓk with $\ell \geq 1$ we have*

$$\phi_M(\ell k) \leq \lceil \ell \rceil \phi_M(k).$$

Proof. Let $\bar{\alpha}$ achieve $\phi_M(\ell k)$. Moreover let $\sum_{i=1}^{\lceil \ell \rceil} \alpha_i = \bar{\alpha}$ such that $\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|_0 = \|\bar{\alpha}\|_0$. We can choose α_i's such that $\|\alpha_i\|_0 \leq k$ since $\lceil \ell \rceil k \geq \ell k$.

Since M is positive semi-definite, for any i, j we have $\alpha_i' M \alpha_i + \alpha_j' M \alpha_j \geq 2 \left|\alpha_i' M \alpha_j\right|$.

Therefore, we have

$$\begin{aligned}\phi_M(\ell k) = \bar{\alpha}' M \bar{\alpha} &= \sum_{i=1}^{\lceil \ell \rceil} \alpha_i' M \alpha_i + \sum_{i=1}^{\lceil \ell \rceil} \sum_{j \neq i} \alpha_i' M \alpha_j \\ &\leq \sum_{i=1}^{\lceil \ell \rceil} \alpha_i' M \alpha_i + \sum_{i=1}^{\lceil \ell \rceil} \sum_{j \neq i} (\alpha_i' M \alpha_i + \alpha_j' M \alpha_j)/2 \\ &= \sum_{i=1}^{\lceil \ell \rceil} \{\alpha_i' M \alpha_i + (\lceil \ell \rceil - 1)\alpha_i' M \alpha_i\} \\ &\leq \lceil \ell \rceil \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 \phi_M(\|\alpha_i\|_0).\end{aligned}$$

Note that $\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 = 1$ and thus $\phi_M(\ell k) \leq \lceil \ell \rceil \max_{i=1,\ldots,\lceil \ell \rceil} \phi_M(\|\alpha_i\|_0) \leq \lceil \ell \rceil \phi_M(k)$. □

The following lemmas characterize the behavior of the maximal sparse eigenvalue for the case of correlated Gaussian regressors. We start by establishing an upper bound on $\phi(k)$ that holds with high probability.

LEMMA 12. *Consider $x_i = \Sigma^{1/2} z_i$, where $z_i \sim N(0, I_p)$, $p \geq n$, and $\sup_{\alpha \in \mathbb{S}_p^k} \alpha' \Sigma \alpha \leq \sigma^2(k)$. Let $\phi(k)$ be the maximal k-sparse eigenvalue of $\mathbb{E}_n[x_i x_i']$, for $k \leq n$. Then with probability converging to one, uniformly in $k \leq n$,*

$$\sqrt{\phi(k)} \lesssim \sigma(k) \left(1 + \sqrt{k/n}\sqrt{\log p}\right).$$

PROOF. By Lemma 11 it suffices to establish the result for $k \leq n/2$. Let Z be the $n \times p$ matrix collecting vectors z_i', $i = 1, \ldots, n$ as rows. Consider the Gaussian process $\mathcal{G}_k : (\alpha, \tilde{\alpha}) \mapsto \tilde{\alpha}' Z \alpha / \sqrt{n}$, where $(\alpha, \tilde{\alpha}) \in \mathbb{S}_p^k \times \mathbb{S}^{n-1}$. Note that

$$\|\mathcal{G}_k\| = \sup_{(\alpha,\tilde{\alpha}) \in \mathbb{S}_p^k \times \mathbb{S}^{n-1}} |\tilde{\alpha}' Z \alpha / \sqrt{n}| = \sup_{\alpha \in \mathbb{S}_p^k} \sqrt{\alpha' \mathbb{E}_n[z_i z_i'] \alpha} = \sqrt{\phi(k)}.$$

Using Borell's concentration inequality for the Gaussian process (see van der Vaart and Wellner [38] Lemma A.2.1) we have that $P\{\|\mathcal{G}_k\| - \text{median}\|\mathcal{G}_k\| > r\} \leq e^{-nr^2/2}$. Also, by classical results on the behavior of the maximal eigenvalues of the Gaussian covariance matrices (see German [13]), as $n \to \infty$, for any $k/n \to \gamma \in [0,1]$, we have that $\lim_{k,n}(\text{median}\|\mathcal{G}_k\| - 1 - \sqrt{k/n}) = 0$. Since k/n lies within $[0,1]$, any sequence k_n/n has convergent subsequence with limit in $[0,1]$. Therefore, we can conclude that, as $n \to \infty$, $\limsup_{k_n,n}(\text{median}\|\mathcal{G}_{k_n}\| - 1 - \sqrt{k_n/n}) \leq 0$. This further implies $\limsup_n \sup_{k \leq n}(\text{median}\|\mathcal{G}_k\| - 1 - \sqrt{k/n}) \leq 0$. Therefore, for any $r_0 > 0$ there exists n_0 large enough such that for all $n \geq n_0$ and all $k \leq n$,

$P\left\{\|\mathcal{G}_k\| > 1+\sqrt{k/n}+r+r_0\right\} \le e^{-nr^2/2}$. There are at most $\binom{p}{k}$ subvectors of z_i containing k elements, so we conclude that for $n \ge n_0$,

$$P\left\{\sup_{\alpha\in\mathbb{S}_p^k}\sqrt{\alpha'\mathbb{E}_n[z_iz_i']\alpha} > 1+\sqrt{k/n}+r_k+r_0\right\} \le \binom{p}{k}e^{-nr_k^2/2}.$$

Summing over $k \le n$ we obtain

$$\sum_{k=1}^n P\left\{\sup_{\alpha\in\mathbb{S}_p^k}\sqrt{\alpha'\mathbb{E}_n[z_iz_i']\alpha} > 1+\sqrt{k/n}+r_k+r_0\right\} \le \sum_{k=1}^n \binom{p}{k}e^{-nr_k^2/2}.$$

Setting $r_k = \sqrt{ck/n\log p}$ for $c > 1$ and using that $\binom{p}{k} \le p^k$, we bound the right side by $\sum_{k=1}^n e^{(1-c)k\log p} \to 0$ as $n \to \infty$. We conclude that with probability converging to one, uniformly for all k: $\sup_{\alpha\in\mathbb{S}_p^k}\sqrt{\alpha'\mathbb{E}_n[z_iz_i']\alpha} \lesssim 1+\sqrt{k/n}\sqrt{\log p}$. Furthermore, since $\sup_{\alpha\in\mathbb{S}_p^k}\alpha'\Sigma\alpha \le \sigma^2(k)$, we conclude that with probability converging to one, uniformly for all k: $\sup_{\alpha\in\mathbb{S}_p^k}\sqrt{\alpha'\mathbb{E}_n[x_ix_i']\alpha} \lesssim \sigma(k)(1+\sqrt{k/n}\sqrt{\log p})$. □

Next, relying on Sudakov's minoration, we show a lower bound on the expectation of the maximum k-sparse eigenvalue. We do not use the lower bound in the analysis, but the result shows that the upper bound is sharp in terms of the rate dependence on k, p, and n.

LEMMA 13. *Consider $x_i = \Sigma^{1/2}z_i$, where $z_i \sim N(0, I_p)$, and $\inf_{\alpha\in\mathbb{S}_p^k}\alpha'\Sigma\alpha \ge \underline{\sigma}^2(k)$. Let $\phi(k)$ be the maximal k-sparse eigenvalue of $\mathbb{E}_n[x_ix_i']$, for $k \le n < p$. Then for any even k we have that:*

$$(1)\ \ \mathrm{E}\left[\sqrt{\phi(k)}\right] \ge \frac{\underline{\sigma}(2k)}{3\sqrt{n}}\sqrt{(k/2)\log(p-k)}\ \text{ and }(2)\ \ \sqrt{\phi(k)} \gtrsim_P \frac{\underline{\sigma}(2k)}{3\sqrt{n}}\sqrt{(k/2)\log(p-k)}.$$

PROOF. Let X be the $n\times p$ matrix collecting vectors x_i', $i = 1,\dots,n$ as rows. Consider the Gaussian process $(\alpha,\tilde{\alpha}) \mapsto \tilde{\alpha}'X\alpha/\sqrt{n}$, where $(\alpha,\tilde{\alpha}) \in \mathbb{S}_p^k \times \mathbb{S}^{n-1}$. Note that $\sqrt{\phi(k)}$ is the supremum of this Gaussian process

$$\sup_{(\alpha,\tilde{\alpha})\in\mathbb{S}_p^k\times\mathbb{S}^{n-1}}|\tilde{\alpha}'X\alpha/\sqrt{n}| = \sup_{\alpha\in\mathbb{S}_p^k}\sqrt{\alpha'\mathbb{E}_n[x_ix_i']\alpha} = \sqrt{\phi(k)}. \tag{A.11}$$

Hence we proceed in three steps: In Step 1, we consider the uncorrelated case and prove the lower bound (1) on the expectation of the supremum using Sudakov's minoration, using a lower bound on a relevant packing number. In Step 2, we derive the lower bound on the packing number. In Step 3, we generalize Step 1 to the correlated case. In Step 4, we prove the lower bound (2) on the supremum itself using Borell's concentration inequality.

Step 1. In this step we consider the case where $\Sigma = I$ and show the result using Sudakov's minoration. By fixing $\tilde{\alpha} = (1, \ldots, 1)'/\sqrt{n} \in \mathbb{S}^{n-1}$, we have $\sqrt{\phi(k)} \geq \sup_{\alpha \in \mathbb{S}^k_p} \mathbb{E}_n[x_i'\alpha] = \sup_{\alpha \in \mathbb{S}^k_p} Z_\alpha$, where $\alpha \mapsto Z_\alpha := \mathbb{E}_n[x_i'\alpha]$ is a Gaussian process on $\mathbb{S}^k_p$. We will bound $E[\sup_{\alpha \in \mathbb{S}^k_p} Z_\alpha]$ from below using Sudakov's minoration.

We consider the standard deviation metric on $\mathbb{S}^k_p$ induced by Z: for any $t, s \in \mathbb{S}^k_p$,

$$d(s,t) = \sqrt{\sigma^2(Z_t - Z_s)} = \sqrt{\mathrm{E}\left[(Z_t - Z_s)^2\right]} = \sqrt{\mathrm{E}[\mathbb{E}_n\left[(x_i'(t-s))^2\right]]} = \|t - s\|/\sqrt{n}.$$

Consider the packing number $D(\epsilon, \mathbb{S}^k_p, d)$, the largest number of disjoint closed balls of radius ϵ with respect to the metric d that can be packed into $\mathbb{S}^k_p$, see [10]. We will bound the packing number from below for $\epsilon = \frac{1}{\sqrt{n}}$. In order to do this we restrict attention to the collection $\mathcal{T}$ of elements $t = (t_1, \ldots, t_p) \in \mathbb{S}^k_p$ such that $t_i = 1/\sqrt{k}$ for exactly k components and $t_i = 0$ in the remaining $p - k$ components. There are $|T| = \binom{p}{k}$ of such elements. Consider any $s, t \in \mathcal{T}$ such that the support of s agrees with the support of t in at most $k/2$ elements. In this case

$$\|s - t\|^2 = \sum_{j=1}^{p} |t_j - s_j|^2 \geq \sum_{\substack{j \in \mathrm{support}(t) \\ \setminus \mathrm{support}(s)}} \frac{1}{k} + \sum_{\substack{j \in \mathrm{support}(s) \\ \setminus \mathrm{support}(t)}} \frac{1}{k} \geq 2\frac{k}{2}\frac{1}{k} = 1. \tag{A.12}$$

Let $\mathcal{P}$ be the set of the maximal cardinality, consisting of elements in $\mathcal{T}$ such that $|\mathrm{support}(t) \setminus \mathrm{support}(s)| \geq k/2$ for every $s, t \in \mathcal{P}$. By the inequality (A.12) we have that $D(1/\sqrt{n}, \mathbb{S}^k_p, d) \geq |\mathcal{P}|$. Furthermore, by Step 2 given below we have that $|\mathcal{P}| \geq (p-k)^{k/2}$.

Using Sudakov's minoration ([12], Theorem 4.1.4), we conclude that

$$\mathrm{E}\Big[\sup_{t \in \mathbb{S}^k_p} Z_t\Big] \geq \sup_{\epsilon > 0} \frac{\epsilon}{3}\sqrt{\log D(\epsilon, \mathbb{S}^k_p, d)} \geq \sqrt{\log D(1/\sqrt{n}, \mathbb{S}^k_p, d)} \geq \frac{1}{3}\sqrt{k \log(p-k)/(2n)},$$

proving the claim of the lemma for the case $\Sigma = I$.

Step 2. In this step we show that $|\mathcal{P}| \geq (p-k)^{k/2}$.

It is convenient to identify every element $t \in \mathcal{T}$ with the set $\mathrm{support}(t)$, where $\mathrm{support}(t) = \{j \in \{1, \ldots, p\} : t_j = 1/\sqrt{k}\}$, which has cardinality k. For any $t \in \mathcal{T}$ let $\mathcal{N}(t) = \{s \in \mathcal{T} : |\mathrm{support}(t) \setminus \mathrm{support}(s)| \leq k/2\}$. By construction we have that $\max_{t \in \mathcal{T}} |\mathcal{N}(t)||\mathcal{P}| \geq |\mathcal{T}|$. Since as shown below $\max_{t \in \mathcal{T}} |\mathcal{N}(t)| \leq K := \binom{k}{k/2}\binom{p-k/2}{k/2}$ for every t, we conclude that $|\mathcal{P}| \geq |\mathcal{T}|/K = \binom{p}{k}/K \geq (p-k)^{k/2}$.

It remains only to show that $|\mathcal{N}(t)| \leq \binom{k}{k/2}\binom{p-k/2}{k/2}$. Consider an arbitrary $t \in \mathcal{T}$. Fix any $k/2$ components of $\mathrm{support}(t)$, and generate elements $s \in \mathcal{N}(t)$ by switching any of the remaining $k/2$ components in $\mathrm{support}(t)$ to any of the possible $p - k/2$ values. This gives us at most $\binom{p-k/2}{k/2}$ such elements $s \in \mathcal{N}(t)$. Next let us repeat this procedure for all other combinations of initial $k/2$ components of support(t), where the number of such combinations is bounded by $\binom{k}{k/2}$. In this way we generate every element $s \in \mathcal{N}(t)$. From the construction we conclude that $|\mathcal{N}(t)| \leq \binom{k}{k/2}\binom{p-k/2}{k/2}$.

Step 3. The case where $\Sigma \neq I$ follows similarly noting that the new metric, $d(s,t) = \sqrt{\sigma^2(Z_t - Z_s)} = \sqrt{\mathrm{E}\left[(Z_t - Z_s)^2\right]}$, satisfies

$$d(s,t) \geq \underline{\sigma}(2k)\|s-t\|/\sqrt{n} \ \text{ since } \ \|s-t\|_0 \leq 2k.$$

Step 4. Using Borell's concentration inequality (see van der Vaart and Wellner [38] Lemma A.2.1) for the supremum of the Gaussian process defined in (A.11), we have $P\{|\sqrt{\phi(k)} - E[\sqrt{\phi(k)}]| > r\} \leq 2e^{-nr^2/2}$, which proves the second claim of the lemma. □

Next we combine the previous lemmas to control the empirical sparse eigenvalues of Examples 1 and 2.

LEMMA 14. *For $k \leq n$, under the design of Example 1 we have $\phi(k) \simeq_p 1 + \sqrt{\frac{k \log p}{n}}$. For $k \leq n$, under the design of Example 2 we have*

$$\phi(k) \lesssim_p \frac{1+|\rho|}{1-|\rho|}\left(1 + \sqrt{\frac{k\log p}{n}}\right) \quad \textit{and} \quad \phi(k) \gtrsim_p \frac{1-|\rho|}{1+|\rho|}\left(1 + \sqrt{\frac{k\log p}{n}}\right).$$

PROOF. Consider Example 1. Let $x_{i,-1}$ denote the ith observation without the first component. Write

$$\mathbb{E}_n\left[x_i x_i'\right] = \begin{bmatrix} 1 & \mathbb{E}_n\left[x_{i,-1}'\right] \\ \mathbb{E}_n\left[x_{i,-1}\right] & 0 \end{bmatrix} + \mathbb{E}_n \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{E}_n\left[x_{i,-1}x_{i,-1}'\right] \end{bmatrix} = M + N.$$

We first bound $\phi_N(k)$. Letting $N_{-1,-1} = \mathbb{E}_n\left[x_{i,-1}x_{i,-1}'\right]$ we have $\phi_N(k) = \phi_{N_{-1,-1}}(k)$. Lemma 12 implies that $\phi_N(k) \lesssim_p 1 + \sqrt{k/n}\sqrt{\log p}$. Lemma 13 bounds $\phi_N(k)$ from below because $\phi_{N_{-1,-1}}(k) \gtrsim_p \sqrt{(k/2n)\log(p-k)}$.

We then bound $\phi_M(k)$. Since $M_{11} = 1$, we have $\phi_M(1) \geq 1$. To produce an upper bound let $w = (a, b')'$ achieve $\phi_M(k)$ where $a \in \mathbb{R}$, $b \in \mathbb{R}^{p-1}$. By definition we have $\|w\| = 1$, $\|w\|_0 \leq k$. Note that $|a| \leq 1$, $\|b\| = \sqrt{1 - |a|^2} \leq 1$, $\|b\|_1 \leq \sqrt{k}\|b\|$. Therefore

$$\begin{aligned} \phi_M(k) = w'Mw &= a^2 + 2ab'\mathbb{E}_n\left[x_{i,-1}\right] \leq 1 + 2b'\mathbb{E}_n\left[x_{i,-1}\right] \\ &\leq 1 + 2\|b\|_1 \|\mathbb{E}_n\left[x_{i,-1}\right]\|_\infty \leq 1 + 2\sqrt{k}\|b\|\|\mathbb{E}_n\left[x_{i,-1}\right]\|_\infty. \end{aligned}$$

Next we bound $\|\mathbb{E}_n[x_{i,-1}]\|_\infty = \max_{j=2,\ldots,p} |\mathbb{E}_n[x_{ij}]|$. Since $\mathbb{E}_n[x_{ij}] \sim N(0,1/n)$ for $j = 2,\ldots,p$, by (4.13) we have $\|\mathbb{E}_n[x_{i,-1}]\|_\infty \lesssim_P \sqrt{(1/n)\log p}$. Therefore we have $\phi_M(k) \lesssim_P 1 + 2\sqrt{k/n}\sqrt{\log p}$.

Finally, we bound ϕ. Note that $\phi(k) = \sup_{\alpha \in \mathbb{S}_p^k} \alpha'(M+N)\alpha = \sup_{\alpha \in \mathbb{S}_p^k} \alpha' M\alpha + \alpha' N\alpha \leq \phi_M(k) + \phi_N(k)$. On the other hand, $\phi(k) \geq 1 \vee \phi_{N_{-1,-1}}(k)$ since the covariates contain an intercept. The result follows by using the bounds derived above.

The proof for the design of Example 2 is similar with the same steps. Since $-1 < \rho < 1$ is fixed, the bounds on the eigenvalues of the population design matrix Σ to apply Lemmas 12 and 13 are given by $\sigma^2(k) = \sup_{\alpha \in \mathbb{S}_p^k} \alpha'\Sigma\alpha \leq (1+|\rho|)/(1-|\rho|)$ and $\underline{\sigma}^2(k) = \inf_{\alpha \in \mathbb{S}_p^k} \alpha'\Sigma\alpha \geq \frac{1}{2}(1-|\rho|)/(1+|\rho|)$. To bound $\phi_M(k)$ comparison theorem (4.15) allows for the same bound as for the uncorrelated design to hold. □

A.4. Maximal Inequalities for a Collection of Empirical Processes.

The main result of this section is Lemma 18, stating a maximal inequality that controls the empirical process uniformly over a collection of classes of functions using class-dependent bounds. We need this lemma, because the standard maximal inequalities applied to the union of function classes yield a single class-independent bound that is too large for our purposes.

We prove the main result by first stating Lemma 15, giving a bound on tail probabilities of a separable sub-Gaussian process, stated in terms of uniform covering numbers. Here we want to explicitly trace the impact of covering numbers on the tail probability, since these covering numbers grow rapidly under increasing parameter dimension. Using the symmetrization approach, we then obtain Lemma 17, giving a bound on tail probabilities of a general separable empirical process, also stated in terms of uniform covering numbers. Finally given a growth rate on the covering numbers, we obtain our final Lemma 18, which we repeatedly employ throughout the paper.

Lemma 15 (Exponential Inequality for Sub-Gaussian Process). *Consider any linear zero-mean separable process $\{\mathbb{G}(f) : f \in \mathcal{F}\}$, whose index set $\mathcal{F}$ includes zero, is equipped with a $L_2(P)$ norm, and has envelope F. Suppose further that the process is sub-Gaussian, namely for each $g \in \mathcal{F} - \mathcal{F}$:*

$$\mathbb{P}\{|\mathbb{G}(g)| > \eta\} \leq 2\exp\left(-\frac{1}{2}\eta^2/D^2\|g\|_{P,2}^2\right) \qquad \text{for any} \quad \eta > 0,$$

with D a positive constant; and suppose that we have the following upper bound for the *uniform L_2 covering numbers for $\mathcal{F}$:*

$$\sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq n(\epsilon, \mathcal{F}, L_2) \text{ for each } \epsilon > 0,$$

where $n(\epsilon, \mathcal{F}, L_2)$ is increasing in $1/\epsilon$, and $\epsilon\sqrt{\log n(\epsilon, \mathcal{F}, L_2)} \to 0$ as $1/\epsilon \to \infty$ and is decreasing in $1/\epsilon$. Then for $K > D$, for some universal constant $c < 30$, $\rho(\mathcal{F}, P) := \sup_{f\in\mathcal{F}} \|f\|_{P,2}/\|F\|_{P,2}$,

$$\mathbb{P}\left\{\frac{\sup_{f\in\mathcal{F}} |\mathbb{G}(f)|}{\|F\|_{P,2}\int_0^{\rho(\mathcal{F},P)/4}\sqrt{\log n(x, \mathcal{F}, L_2)}dx} > cK\right\} \leq \int_0^{\rho(\mathcal{F},P)/2} \epsilon^{-1} n(\epsilon, \mathcal{F}, L_2)^{-\{(K/D)^2-1\}} d\epsilon.$$

The result of Lemma 15 is similar in spirit to the result of Ledoux and Talagrand [23], page 302, on tail probabilities of a process stated in terms of Orlicz-norm covering numbers. However, Lemma 15 gives a tail probability stated in terms of the uniform L_2 covering numbers. The reason is that in our context estimates of the uniform L_2 covering numbers for common function classes are more readily available than the estimates the Orlicz-norm covering numbers.

In order to prove a bound on tail probabilities of a general separable empirical process, we need to go through a symmetrization argument. Since we use data-dependent threshold, we need an appropriate extension of the classical symmetrization lemma to allow for this. Let us call a threshold function $x : \mathbb{R}^n \mapsto \mathbb{R}$ k-sub-exchangeable if for any $v, w \in \mathbb{R}^n$ and any vectors $\tilde{v}, \tilde{w}$ created by the pairwise exchange of the components in v with components in w, we have that $x(\tilde{v}) \vee x(\tilde{w}) \geq [x(v) \vee x(w)]/k$. Several functions satisfy this property, in particular $x(v) = \|v\|$ with $k = \sqrt{2}$ and constant functions with $k = 1$. The following result generalizes the standard symmetrization lemma for probabilities (Lemma 2.3.7 of [38]) to the case of a random threshold x that is sub-exchangeable.

LEMMA 16 (Symmetrization with Data-dependent Threshold). *Consider arbitrary independent stochastic processes $Z_1, \ldots, Z_n$ and arbitrary functions $\mu_1, \ldots, \mu_n : \mathcal{F} \mapsto \mathbb{R}$. Let $x(Z) = x(Z_1, \ldots, Z_n)$ be a k-sub-exchangeable random variable and for any $\tau \in (0, 1)$ let q_τ denote the τ quantile of $x(Z)$, $\bar{p}_\tau := P(x(Z) \leq q_\tau) \geq \tau$, and $p_\tau := P(x(Z) < q_\tau) \leq \tau$. We have*

$$P\left(\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}} > x_0 \vee x(Z)\right) \leq \frac{4}{\bar{p}_\tau} P\left(\left\|\sum_{i=1}^n \varepsilon_i (Z_i - \mu_i)\right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z)}{4k}\right) + p_\tau$$

where x_0 is a constant such that $\inf_{f\in\mathcal{F}} P\left(|\sum_{i=1}^n Z_i(f)| \leq \frac{x_0}{2}\right) \geq 1 - \frac{\bar{p}_\tau}{2}$.

Note that we can recover the classical symmetrization lemma where threshold is fixed by setting $k = 1$, $\bar{p}_\tau = 1$, and $p_\tau = 0$. The next lemma follows from combining the previous two lemmas.

LEMMA 17 (Exponential inequality for separable empirical process). *Consider a separable empirical process $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n\{f(Z_i) - \mathrm{E}[f(Z_i)]\}$, where $Z_1, \ldots, Z_n$ is an underlying i.i.d. data sequence. Let $K > 1$ and $\tau \in (0,1)$ be constants, and $e_n(\mathcal{F}, \mathbb{P}_n) = e_n(\mathcal{F}, Z_1, \ldots, Z_n)$ be a k-sub-exchangeable random variable, such that*

$$\|F\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F},\mathbb{P}_n)/4}\sqrt{\log n(\epsilon, \mathcal{F}, L_2)}d\epsilon \leq e_n(\mathcal{F}, \mathbb{P}_n) \text{ and } \sup_{f\in\mathcal{F}} var_{\mathbb{P}} f \leq \frac{\tau}{2}(4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2$$

for the same constant $c > 0$ as in Lemma 15, then

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)| \geq 4kcKe_n(\mathcal{F}, \mathbb{P}_n)\right\} \leq \frac{4}{\tau}\mathrm{E}_{\mathbb{P}}\left(\left[\int_0^{\rho(\mathcal{F},\mathbb{P}_n)/2}\epsilon^{-1}n(\epsilon, \mathcal{F}, L_2)^{-\{K^2-1\}}d\epsilon\right] \wedge 1\right) + \tau.$$

Finally, our main result in this section is as follows.

LEMMA 18 (Maximal Inequality for a Collection of Empirical Processes). *Consider a collection of separable empirical processes $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n\{f(Z_i) - \mathrm{E}[f(Z_i)]\}$, where $Z_1, \ldots, Z_n$ is an underlying i.i.d. data sequence, defined over function classes $\mathcal{F}_m, m = 1, \ldots, n$ with envelopes $F_m = \sup_{f\in\mathcal{F}_m}|f(x)|, m = 1, \ldots, n$, and with upper bounds on the uniform covering numbers of $\mathcal{F}_m$ given for all m by*

$$n(\epsilon, \mathcal{F}_m, L_2) = (n \vee p)^m(\kappa/\epsilon)^{vm}, \ 0 < \epsilon < 1,$$

with some constants $\kappa > 1$ and $v > 1$. For a constant $C := (1 + \sqrt{2v})/4$ set

$$e_n(\mathcal{F}_m, \mathbb{P}_n) = C\sqrt{m\log(n \vee p)}\max\left\{\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P},2}, \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\right\}.$$

Then, for any $\delta \in (0,1)$, there is a large enough constant $K \geq \sqrt{2/\delta}$, for n sufficiently large, the inequality

$$\sup_{f\in\mathcal{F}_m}|\mathbb{G}_n(f)| \leq 4\sqrt{2}cKe_n(\mathcal{F}_m, \mathbb{P}_n), \text{ for all } m \leq n,$$

holds with probability at least $1 - \delta$, where the constant c is the same as in Lemma 15.

Now we prove Lemmas 15, 16, 17, and 18.

Proof of Lemma 15. The proof follows by specializing arguments given van der Vaart [36], page 286, to the sub-Gaussian processes and also tracing out the bounds on tail probabilities in full detail.

Step 1. There exists a sequence of nested partitions of $\mathcal{F}$, $\{(\mathcal{F}_{qi}, i = 1, \dots, N_q), q = q_0, q_0 + 1, \dots\}$ where the q-th partition consists of sets of $L_2(P)$ radius at most $\|F\|_{P,2}2^{-q}$, where q_0 is the largest positive integer such that $2^{-q_0} \leq \rho(\mathcal{F}, P)/4$ so that $q_0 \geq 2$. The existence of such partition follows from a standard argument, e.g. van der Vaart [36], page 286, which we repeat here: To construct the q-th partition, cover $\mathcal{F}$ with at most $n_q = n(2^{-q}, \mathcal{F}, L_2)$ balls of $L_2(P)$ radius $\|F\|_{P,2}2^{-q}$ and replace these by the same number of disjoint sets. If the sequence of partitions does not yet consist of successive refinements, then replace the partition at stage q by the set of all intersections of the form $\cap_{j=q_0}^{q} \mathcal{F}_{ji}$. This gives partition into at most $N_q = n_{q_0} \cdots n_q$ sets, so that $\log N_q = \sum_{j=q_0}^{q} \log n_j$.

Let f_{qi} be an arbitrary point of $\mathcal{F}_{qi}$. Set $\pi_q(f) = f_{qi}$ if $f \in \mathcal{F}_{qi}$. By separability of the process, we can replace $\mathcal{F}$ by $\cup_{q,i} f_{qi}$, since the supremum norm of the process can be computed by taking this set only. In this case, we can decompose $f - \pi_{q_0}(f) = \sum_{q=q_0+1}^{\infty}(\pi_q(f) - \pi_{q-1}(f))$. Hence by linearity

$$\mathbb{G}(f) - \mathbb{G}(\pi_{q_0}(f)) = \sum_{q=q_0+1}^{\infty} \mathbb{G}(\pi_q(f)) - \mathbb{G}(\pi_{q-1}(f)) = \sum_{q=q_0+1}^{\infty} \mathbb{G}(\pi_q(f) - \pi_{q-1}(f)),$$

and $|\mathbb{G}(f)| \leq \sum_{q=q_0+1}^{\infty} \max_f |\mathbb{G}(\pi_q(f) - \pi_{q-1}(f))| + \max_f |\mathbb{G}(\pi_{q_0}(f))|$. Thus

$$\mathbb{P}\Big\{ \sup_{f \in \mathcal{F}} |\mathbb{G}(f)| > \sum_{q=q_0}^{\infty} \eta_q \Big\} \leq \sum_{q=q_0+1}^{\infty} \mathbb{P}\left\{ \max_f |\mathbb{G}(\pi_q(f) - \pi_{q-1}(f))| > \eta_q \right\} + \mathbb{P}\left\{ \max_f |\mathbb{G}(\pi_{q_0}(f))| > \eta_{q_0} \right\},$$

for constants η_q chosen below.

Step 2. By construction of the partition sets

$$\|\pi_q(f) - \pi_{q-1}(f)\|_{P,2} \leq 2\|F\|_{P,2}2^{-(q-1)} \leq 4\|F\|_{P,2}2^{-q}, \text{ for } q \geq q_0 + 1.$$

Setting $\eta_q = 8K\|F\|_{P,2}2^{-q}\sqrt{\log N_q}$, using sub-Gaussianity, setting $K > D$, using that $2\log N_q \geq \log N_q N_{q-1} \geq \log n_q$, using that $q \mapsto \log n_q$ is increasing in q, and $2^{-q_0} \leq$

$\rho(\mathcal{F}, P)/4$, we obtain

$$\begin{aligned}
\sum_{q=q_0+1}^{\infty} \mathbb{P}\left\{\max_f |\mathbb{G}(\pi_q(f) - \pi_{q-1}(f))| > \eta_q\right\} &\leq \sum_{q=q_0+1}^{\infty} N_q N_{q-1} 2 \exp\left(-\eta_q^2/(4D\|F\|_{P,2}2^{-q})^2\right) \\
&\leq \sum_{q=q_0+1}^{\infty} N_q N_{q-1} 2 \exp\left(-(K/D)^2 2 \log N_q\right) \\
&\leq \sum_{q=q_0+1}^{\infty} 2 \exp\left(-\{(K/D)^2 - 1\} \log(N_q N_{q-1})\right) \\
&\leq \sum_{q=q_0+1}^{\infty} 2 \exp\left(-\{(K/D)^2 - 1\} \log n_q\right) \\
&\leq \int_{q_0}^{\infty} 2 \exp\left(-\{(K/D)^2 - 1\} \log n_q\right) dq \\
&= \int_0^{\rho(\mathcal{F},P)/4} (x \ln 2)^{-1} 2 n(x, \mathcal{F}, L_2(P))^{-\{(K/D)^2-1\}} dx.
\end{aligned}$$

By Jensen's inequality $\sqrt{\log N_q} \leq a_q := \sum_{j=q_0}^{q} \sqrt{\log n_q}$, so that

$$\sum_{q=q_0+1}^{\infty} \eta_q \leq 8 \sum_{q=q_0+1}^{\infty} K\|F\|_{P,2} 2^{-q} a_q.$$

Letting $b_q = 2 \cdot 2^{-q}$, noting $a_{q+1} - a_q = \sqrt{\log n_{q+1}}$ and $b_{q+1} - b_q = -2^{-q}$, we get using summation by parts

$$\begin{aligned}
\sum_{q=q_0+1}^{\infty} 2^{-q} a_q &= -\sum_{q=q_0+1}^{\infty} (b_{q+1} - b_q) a_q = -\left(a_q b_q|_{q_0+1}^{\infty} - \sum_{q=q_0+1}^{\infty} b_{q+1}(a_{q+1} - a_q)\right) \\
&= \left(2 \cdot 2^{-(q_0+1)} \sqrt{\log n_{q_0+1}} + \sum_{q=q_0+1}^{\infty} 2 \cdot 2^{-(q+1)} \sqrt{\log n_{q+1}}\right) = 2 \sum_{q=q_0+1}^{\infty} 2^{-q} \sqrt{\log n_q},
\end{aligned}$$

where we use the assumption that $2^{-q}\sqrt{\log n_q} \to 0$ as $q \to \infty$, so that $-a_q b_q|_{q_0+1}^{\infty} = 2 \cdot 2^{-(q_0+1)}\sqrt{\log n_{q_0+1}}$. Using that $2^{-q}\sqrt{\log n_q}$ is decreasing in q by assumption,

$$2 \sum_{q=q_0+1}^{\infty} 2^{-q} \sqrt{\log n_q} \leq 2 \int_{q_0}^{\infty} 2^{-q} \sqrt{\log n(2^{-q}, \mathcal{F}, L_2(P))} dq.$$

Using a change of variables and that $2^{-q_0} \leq \rho(\mathcal{F}, P)/4$, we finally conclude that

$$\sum_{q=q_0+1}^{\infty} \eta_q \leq K\|F\|_{P,2} \frac{16}{\log 2} \int_0^{\rho(\mathcal{F},P)/4} \sqrt{\log n(x, \mathcal{F}, L_2(P))} dx.$$

Step 3. Letting $\eta_{q_0} = K\|F\|_{P,2}\rho(\mathcal{F}, P)\sqrt{2 \log N_{q_0}}$, recalling that $N_{q_0} = n_{q_0}$, using $\|\pi_{q_0}(f)\|_{P,2} \leq$

$\|F\|_{P,2}$ and sub-Gaussianity, we conclude

$$\mathbb{P}\Big\{ \max_{f} |\mathbb{G}(\pi_{q_0}(f))| > \eta_{q_0} \Big\} \leq n_q 2 \exp\Big(-(K/D)^2 \log n_q\Big) \leq 2 \exp\Big(-\{(K/D)^2 - 1\} \log n_q\Big)$$
$$\leq \int_{q_0-1}^{q_0} 2 \exp\Big(-\{(K/D)^2 - 1\} \log n_q\Big)\, dq = \int_{\rho(\mathcal{F},P)/4}^{\rho(\mathcal{F},P)/2} (x \ln 2)^{-1} 2n(x, \mathcal{F}, L_2(P))^{-\{(K/D)^2-1\}} dx.$$

Also, since $n_{q_0} = n(2^{-q_0}, \mathcal{F}, P)$, $2^{-q_0} \leq \rho(\mathcal{F}, P)/4$, and $n(x, \mathcal{F}, P)$ is increasing in $1/x$, we obtain:

$$\eta_{q_0} = 4K\|F\|_{P,2}[\rho(\mathcal{F}, P)/4]\sqrt{2 \log n(2^{-q_0}, \mathcal{F}, P)} \leq 4\sqrt{2}K\|F\|_{P,2} \int_0^{\rho(\mathcal{F},P)/4} \sqrt{\log n(x, \mathcal{F}, P)}dx.$$

Step 4. Finally, adding the bounds on tail probabilities from Steps 2 and 3 we obtain the tail bound stated in the main text. Further, adding bounds on η_q from Steps 2 and 3, and using $c = 16/\log 2 + 4\sqrt{2} < 30$, we obtain

$$\sum_{q=q_0}^{\infty} \eta_q \leq cK\|F\|_{P,2} \int_0^{\rho(\mathcal{F},P)/4} \sqrt{\log n(x, \mathcal{F}, L_2(P))}dx.$$

□

PROOF OF LEMMA 16. The proof proceeds analogously to the proof of Lemma 2.3.7 (page 112) in [38] with the necessary adjustments. Letting q_τ be the τ quantile of $x(Z)$ we have

$$P\left\{ \left\| \sum_{i=1}^{n} Z_i \right\|_{\mathcal{F}} > x_0 \vee x(Z) \right\} \leq P\left\{ x(Z) \geq q_\tau, \left\| \sum_{i=1}^{n} Z_i \right\|_{\mathcal{F}} > x_0 \vee x(Z) \right\} + P\{x(Z) < q_\tau\}.$$

Next we bound the first term of the expression above. Let $Y = (Y_1, \ldots, Y_n)$ be an independent copy of $Z = (Z_1, \ldots, Z_n)$, suitably defined on a product space. Fix a realization of Z such that $x(Z) \geq q_\tau$ and $\|\sum_{i=1}^n Z_i\|_{\mathcal{F}} > x_0 \vee x(Z)$. Therefore $\exists f_Z \in \mathcal{F}$ such that $|\sum_{i=1}^n Z_i(f_Z)| > x_0 \vee x(Z)$. Conditional on such Z and using the triangular inequality we have that

$$\begin{array}{rcl} P_Y\left\{x(Y) \leq q_\tau, |\sum_{i=1}^n Y_i(f_Z)| \leq \frac{x_0}{2}\right\} & \leq & P_Y\left\{|\sum_{i=1}^n (Y_i - Z_i)(f_Z)| > \frac{x_0 \vee x(Z) \vee x(Y)}{2}\right\} \\ & \leq & P_Y\left\{\|\sum_{i=1}^n (Y_i - Z_i)\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2}\right\}. \end{array}$$

By definition of x_0 we have $\inf_{f \in \mathcal{F}} P\left\{|\sum_{i=1}^n Y_i(f)| \leq \frac{x_0}{2}\right\} \geq 1 - \bar{p}_\tau/2$. Since $P_Y\{x(Y) \leq q_\tau\} = \bar{p}_\tau$, by Bonferroni inequality we have that the left hand side is bounded from below by $\bar{p}_\tau - \bar{p}_\tau/2 = \bar{p}_\tau/2$. Therefore, over the set $\{Z : x(Z) \geq q_\tau, \|\sum_{i=1}^n Z_i\|_{\mathcal{F}} > x_0 \vee x(Z)\}$ we have

$$\frac{\bar{p}_\tau}{2} \leq P_Y\left\{ \left\| \sum_{i=1}^{n} (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2} \right\}.$$

Integrating over Z we obtain

$$\frac{\bar{p}_\tau}{2} P\left\{ x(Z) \geq q_\tau, \left\| \sum_{i=1}^{n} Z_i \right\|_{\mathcal{F}} > x_0 \vee x(Z) \right\} \leq P_Z P_Y \left\{ \left\| \sum_{i=1}^{n} (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2} \right\}.$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ be an independent sequence of Rademacher random variables. Given $\varepsilon_1, \ldots, \varepsilon_n$, set $(\tilde{Y}_i = Y_i, \tilde{Z}_i = Z_i)$ if $\varepsilon_i = 1$ and $(\tilde{Y}_i = Z_i, \tilde{Z}_i = Y_i)$ if $\varepsilon_i = -1$. That is, we create vectors $\tilde{Y}$ and $\tilde{Z}$ by pairwise exchanging their components; by construction, conditional on each $\varepsilon_1, \ldots, \varepsilon_n$, $(\tilde{Y}, \tilde{Z})$ has the same distribution as (Y, Z). Therefore,

$$P_Z P_Y \left\{ \left\| \sum_{i=1}^{n} (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2} \right\} = E_\varepsilon P_Z P_Y \left\{ \left\| \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{Z}_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(\tilde{Z}) \vee x(\tilde{Y})}{2} \right\}.$$

By $x(\cdot)$ being k-sub-exchangeable, and since $\varepsilon_i(Y_i - Z_i) = (\tilde{Y}_i - \tilde{Z}_i)$, we have that

$$E_\varepsilon P_Z P_Y \left\{ \left\| \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{Z}_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(\tilde{Z}) \vee x(\tilde{Y})}{2} \right\} \leq E_\varepsilon P_Z P_Y \left\{ \left\| \sum_{i=1}^{n} \varepsilon_i (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2k} \right\}.$$

By the triangular inequality and removing $x(Y)$ or $x(Z)$, the latter is bounded by

$$P\left\{ \left\| \sum_{i=1}^{n} \varepsilon_i (Y_i - \mu_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Y)}{4k} \right\} + P\left\{ \left\| \sum_{i=1}^{n} \varepsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z)}{4k} \right\}.$$

□

Proof of Lemma 17. We would like to apply exponential inequalities to the general separable empirical process $\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^{n} \{f(Z_i) - \mathrm{E}[f(Z_i)]\}$, which is not sub-Gaussian; here $Z_1, \ldots, Z_n$ is an underlying i.i.d. data sequence. To achieve this we use the standard symmetrization approach. Indeed, we first introduce the symmetrized process $\mathbb{G}_n^o(f) = n^{-1/2} \sum_{i=1}^{n} \{\varepsilon_i f(Z_i)\}$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables, i.e., $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$, which are independent of $Z_1, \ldots, Z_n$. Then the tail probabilities of the general empirical process are bounded by the tail probabilities of the symmetrized process using the symmetrization lemma recalled below. Further, we know that by Hoeffding inequality the symmetrized process is sub-Gaussian conditional on $Z_1, \ldots, Z_n$ with respect to the $L_2(\mathbb{P}_n)$ norm, where $\mathbb{P}_n$ is the empirical measure, and this delivers the result.

By the Chebyshev's inequality and the assumption on $e_n(\mathcal{F}, \mathbb{P}_n)$ we have for the constant τ fixed in the statement of the lemma

$$P(|\mathbb{G}_n(f)| > 4kcKe_n(\mathcal{F}, \mathbb{P}_n)\ ) \leq \frac{\sup_f var_{\mathbb{P}} \mathbb{G}_n(f)}{(4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2} = \frac{\sup_{f \in \mathcal{F}} \mathrm{var}_{\mathbb{P}} f}{(4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2} \leq \tau/2.$$

Therefore, by the symmetrization Lemma 16 we obtain

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)| > 4kcKe_n(\mathcal{F},\mathbb{P}_n)\right\} \leq \frac{4}{\tau}\mathbb{P}\left\{\sup_{f\in\mathcal{F}}|\mathbb{G}_n^{o}(f)| > cKe_n(\mathcal{F},\mathbb{P}_n)\right\} + \tau.$$

We then condition on the values of $Z_1, \ldots, Z_n$, denoting the conditional probability measure as $\mathbb{P}_\varepsilon$. Conditional on $Z_1, \ldots, Z_n$, by the Hoeffding inequality the symmetrized process $\mathbb{G}_n^o$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$ norm, namely for $g \in \mathcal{F} - \mathcal{F}$

$$\mathbb{P}_\varepsilon\{\mathbb{G}_n^o(g) > x\} \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{\|g\|_{\mathbb{P}_n,2}^2}\right).$$

Hence by Lemma 15 with $D = 1$, we can bound

$$\mathbb{P}_\varepsilon\left\{\sup_{f\in\mathcal{F}}|\mathbb{G}_n^o(f)| \geq cKe_n(\mathcal{F},\mathbb{P}_n)\right\} \leq \left[\int_0^{\rho(\mathcal{F},\mathbb{P}_n)/2} \epsilon^{-1}n(\epsilon,\mathcal{F},L_2)^{-\{K^2-1\}}d\epsilon\right] \wedge 1.$$

The result follows from taking the expectation over $Z_1, \ldots, Z_n$. □

Proof of Lemma 18. The proof proceeds in two steps, with the first step containing the main argument and the second step containing some auxiliary calculations.

Step 1. In this step we prove the main result. First, we observe that the bound $\epsilon \mapsto n(\epsilon, \mathcal{F}_m, L_2)$ satisfies the monotonicity hypotheses of Lemma 17 uniformly in $m \leq n$.

Second, recall that $e_n(\mathcal{F}_m,\mathbb{P}_n) := C\sqrt{m\log(n\vee p)}\max\{\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P},2}, \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\}$ for $C = (1+\sqrt{2v})/4$. Note that $\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}$ is $\sqrt{2}$-sub-exchangeable and $\rho(\mathcal{F}_m,\mathbb{P}_n) := \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}/\|F_m\|_{\mathbb{P}_n,2} \geq 1/\sqrt{n}$ by Step 2 below. Thus, uniformly in $m \leq n$:

$$\begin{aligned}
\|F_m\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F}_m,\mathbb{P}_n)/4} & \sqrt{\log n(\epsilon,\mathcal{F},L_2)}d\epsilon \leq \|F_m\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F}_m,\mathbb{P}_n)/4} \sqrt{m\log(n\vee p) + vm\log(\kappa/\epsilon)}d\epsilon \\
& \leq \ (1/4)\sqrt{m\log(n\vee p)}\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2} + \|F_m\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F}_m,\mathbb{P}_n)/4}\sqrt{vm\log(\kappa/\epsilon)}d\epsilon \\
& \leq \ \sqrt{m\log(n\vee p)}\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\left(1+\sqrt{2v}\right)/4 \\
& \leq \ e_n(\mathcal{F}_m,\mathbb{P}_n),
\end{aligned}$$

which follows by $\int_0^\rho \sqrt{\log(\kappa/\epsilon)}d\epsilon \leq \left(\int_0^\rho 1d\epsilon\right)^{1/2}\left(\int_0^\rho \log(\kappa/\epsilon)d\epsilon\right)^{1/2} \leq \rho\sqrt{2\log n}$, for $1/\sqrt{n} \leq \rho \leq 1$ and $\kappa < n$ for n sufficiently large.

Third, set $K := \sqrt{2/\delta} > 1$ so that $B(K) := (K^2-1) = 2/\delta$, and let $\tau_m = \delta/(2m\log(n\vee p))$. Recall that $4\sqrt{2}cC > 1$ where $1 < c < 30$ is defined in Lemma 15. Note that for any $m \leq n$ and $f \in \mathcal{F}_m$, we have by Chebyshev inequality

$$P(|\mathbb{G}_n(f)| > 4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n)\,) \leq \frac{\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P},2}^2}{(4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n))^2} \leq \frac{\delta/4}{(4\sqrt{2}cC)^2m\log(n\vee p)} \leq \tau_m/2.$$

Using Lemma 17 with our choice of τ_m, $m \leq n$, $\kappa > 1$, $\upsilon > 1$, and $\rho(\mathcal{F}_m, \mathbb{P}_n) \leq 1$, we obtain

$$\begin{aligned}
\mathbb{P}\Big\{ \sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| > 4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n), \quad \exists \quad m \leq n \Big\} &\leq \sum_{m=1}^{n} \mathbb{P}\left\{ \sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| > 4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n) \right\} \\
&\leq 4\sum_{m=1}^{n} \left[\frac{(n\vee p)^{-B(K)m}}{\tau_m} \int_0^{1/2} (\kappa/\epsilon)^{-\upsilon B(K)m+1} d\epsilon + \tau_m \right] \\
&\leq 4\sum_{m=1}^{n} \frac{(n\vee p)^{-B(K)m}}{\tau_m} \frac{1}{\upsilon B(K)m} + \sum_{m=1}^{n} \tau_m \\
&< 8(n\vee p)^{-B(K)} \log(n\vee p) + \frac{\delta}{2} \frac{(1+\log n)}{\log(n\vee p)} < \delta
\end{aligned}$$

by our choice of $B(K)$ and n sufficiently large.

Step 2. In this step we perform some auxiliary calculations.

To establish that $\sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n,2}$ is $\sqrt{2}$-sub-exchangeable, let $\tilde{Z}, \tilde{Y}$ be created by pairwise exchanging any components of Z and Y. Then $\sqrt{2}\left(\sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n(\tilde{Z}),2} \vee \sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n(\tilde{Y}),2}\right) \geq \left\{\sup_{f\in\mathcal{F}_m} \|f\|^2_{\mathbb{P}_n(\tilde{Z}),2} + \sup_{f\in\mathcal{F}_m} \|f\|^2_{\mathbb{P}_n(\tilde{Y}),2}\right\}^{1/2} \geq \left\{\sup_{f\in\mathcal{F}_m} \mathbb{E}_n\left[f(\tilde{Z}_i)^2\right] + \mathbb{E}_n\left[f(\tilde{Y}_i)^2\right]\right\}^{1/2} = \left\{\sup_{f\in\mathcal{F}_m} \mathbb{E}_n\left[f(Z_i)^2\right] + \mathbb{E}_n\left[f(Y_i)^2\right]\right\}^{1/2} \geq \left\{\sup_{f\in\mathcal{F}_m} \|f\|^2_{\mathbb{P}_n(Z),2} \vee \sup_{f\in\mathcal{F}_m} \|f\|^2_{\mathbb{P}_n(Y),2}\right\}^{1/2} = \sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n(Z),2} \vee \sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n(Y),2}$.

Next we show that $\rho(\mathcal{F}_m, \mathbb{P}_n) := \sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n,2}/\|F_m\|_{\mathbb{P}_n,2} \geq 1/\sqrt{n}$ for $m \leq n$. The latter follows from $\mathbb{E}_n\left[F_m^2\right] = \mathbb{E}_n[\sup_{f\in\mathcal{F}_m} |f(Z_i)|^2] \leq \sup_{i\leq n} \sup_{f\in\mathcal{F}_m} |f(Z_i)|^2$, and from $\sup_{f\in\mathcal{F}_m} \mathbb{E}_n[|f(Z_i)|^2] \geq \sup_{f\in\mathcal{F}_m} \sup_{i\leq n} |f(Z_i)|^2/n$. □

ACKNOWLEDGEMENTS

We would like to thank Arun Chandrasekhar, Moshe Cohen, Ye Luo, and Pierre-Andre Maugis for thorough proof-reading of several versions of this paper and their detailed comments that helped to considerably improve the paper.

REFERENCES

[1] H. Akaike (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, AC-19, 716-723.

[2] R. J. Barro and J.-W. Lee (1994). Data set for a panel of 139 countries, Discussion paper, NBER, http://www.nber.org/pub/barro.lee.

[3] R. J. Barro and X. Sala-i-Martin (1995). Economic Growth. McGraw-Hill, New York.

[4] A. Belloni and V. Chernozhukov (2008). On the Computational Complexity of MCMC-based Estimators in Large Samples, forthcoming in the *Annals of Statistics*.

[5] A. Belloni and V. Chernozhukov (2008). Conditional Quantile Processes under Increasing Dimension, Duke Technical Report.

[6] D. Bertsimas and J. Tsitsiklis (1997). Introduction to Linear Optimization, Athena Scientific.

[7] M. Buchinsky (1994). Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression Econometrica, Vol. 62, No. 2 (Mar.), pp. 405–458.

[8] E. Candes and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. Volume 35, Number 6, 2313-2351.

[9] V. Chernozhukov (2005). Extremal quantile regression. Ann. Statist. 33, no. 2, 806–839.

[10] R. Dudley (2000). Uniform Cental Limit Theorems, Cambridge Studies in advanced mathematics.

[11] J. Fan and J. Lv (2008). Sure Independence Screening for Ultra-High Dimensional Feature Space, Journal of the Royal Statistical Society Series B, 70, 849-911.

[12] X. Fernique (1997). Fonctions aléatoires gaussiennes, ecteurs aléatoires gaussiens. Publications du Centre de Recherches Mathématiques, Montréal., Ann. Probab. Volume 8, Number 2, 252-261.

[13] S. German (1980). A Limit Theorem for the Norm of Random Matrices, Ann. Probab. Volume 8, Number 2, 252-261.

[14] C. Gutenbrunner and J. Jurečková (1992). Regression Rank Scores and Regression Quantiles The Annals of Statistics, Vol. 20, No. 1 (Mar.), pp. 305-330.

[15] X. He and Q.-M. Shao (2000). On Parameters of Increasing Dimenions, Journal of Multivariate Analysis **73**, 120-135.

[16] P. J. Huber (1993). Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799–821.

[17] K. Knight (1998). Limiting distributions for L_1 regression estimators under general conditions, Annals of Statistics, 26, no. 2, 755–770.

[18] K. Knight and Fu, W. J. (2000). Asymptotics for lasso-type estimators. Ann. Statist. 28 1356-1378.

[19] R. Koenker (2005). Quantile regression, Econometric Society Monographs, Cambridge University Press.

[20] R. Koenker and G. Basset (1978). Regression Quantiles, Econometrica, Vol. 46, No. 1, January, 33–50.

[21] R. Koenker and J. Machado (1999). Goodness of fit and related inference process for quantile regression Journal of the American Statistical Association, 94, 1296–1310.

[22] P.-S. Laplace (1818). *Théorie analytique des probabilités.* Éditions Jacques Gabay (1995), Paris.

[23] M. Ledoux and M. Talagrand (1991). Probability in Banach Spaces (Isoperimetry and processes). Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag.

[24] R. Levine and D. Renelt (1992). A Sensitivity Analysis of Cross-Country Growth Regressions, The American Economic Review, Vol. 82, No. 4, pp. 942-963.

[25] N. Meinshausen and P. Buhlmann (2006). High dimensional graphs and variable selection with the Lasso. Ann. Statist. 34 1436-1462.

[26] N. Meinshausen and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data, The Annals of Statistics, Vol. 37, No. 1, 246270.

[27] Y. Nesterov and A. Nemirovskii (1993). Interior-Point Polynomial Algorithms in Convex Programming, vol 13, SIAM Studies in Applied Mathematics, Philadelphia.

[28] S. Portnoy (1991). Asymptotic behavior of regression quantiles in nonstationary, dependent cases.

J. Multivariate Anal. 38, no. 1, 100–113.

[29] S. PORTNOY AND R. KOENKER (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Statist. Sci. Volume 12, Number 4, 279-300.

[30] B. RECHT, M. FAZEL, AND P. A. PARRILO (2007). Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization, Arxiv preprint arXiv:0706.4138, submitted.

[31] X. X. SALA-I-MARTIN(1997). I Just Ran Two Million Regressions, The American Economic Review, Vol. 87, No. 2, pp. 178-183.

[32] G. SCHWARZ (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

[33] A. N. SHIRYAEV (1995). Probability, Springer.

[34] W. F. STOUT (1974). Almost sure convergence. Probability and Mathematical Statistics, Academic Press.

[35] R. TIBSHIRANI (1996). Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B, 58, 267-288.

[36] A. W. VAN DER VAART (1998). Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics.

[37] S. A. VAN DER GEER (2008). High-dimensional generalized linear models and the lasso, Annals of Statistics, Vol. 36, No. 2, 61–645.

[38] A. W. VAN DER VAART AND J. A. WELLNER (1996). Weak Convergence and Empirical Processes, Springer Series in Statistics.

[39] C.-H. ZHANG AND J. HUANG (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. Volume 36, Number 4, 1567-1594.

ALEXANDRE BELLONI
DUKE UNIVERSITY .
FUQUA SCHOOL OF BUSINESS
1 TOWERVIEW DRIVE
DURHAM, NC 27708-0120
PO BOX 90120
E-MAIL: abn5@duke.edu

VICTOR CHERNOZHUKOV
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
DEPARTMENT OF ECONOMICS AND OPERATIONS RESEARCH CENTER
50 MEMORIAL DRIVE
ROOM E52-262F
CAMBRIDGE, MA 02142
E-MAIL: vchern@mit.edu